Engineering of Allosteric Transcription Factors and Their Use for Metabolic Pathway Evolution

Engineering of allosteric transcription factors and their use for metabolic pathway evolution

A dissertation presented

by

Noah David Taylor

to the

Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological and Biomedical Sciences

Harvard University

Cambridge, Massachusetts

October, 2015

Dissertation Advisor: Professor George M. Church Noah David Taylor

Engineering of allosteric transcription factors and their use for metabolic pathway evolution

Abstract

Microbial metabolic production is an attractive alternative to traditional chemical synthesis for a wide array of commercially relevant molecules. Coaxing microbes to produce a target chemical efficiently often requires substantial modification of host cell metabolism, which necessitates searching a vast genetic space of enzyme genes and expression levels. Millions of pathway designs can now be built, but identifying the most productive cells remains low throughput. The ability to detect and report on the presence of any arbitrary target molecule within individual cells would transform the field of metabolic engineering.

To this end, we developed strains of *E. coli* that survive an antibiotic challenge only in the presence of a specific small molecule, by regulating resistance gene expression via transcription factors responsive to sugars, alkanes, macrolides, flavonoids, vitamins or other molecules. Using two of these whole cell biosensors, responsive to glucaric acid or naringenin, we evolved respective biosynthetic pathways for each compound toward higher production. We used oligonucleotide-mediated genomic editing to simultaneously target up to 20 enzyme genes for expression modulation or knockout, creating billions of unique strains. Demonstrating the first example of iterative, whole-pathway engineering via a metabolite biosensor, we discovered *E. coli* strains that had increased production of naringenin by 36 times, or glucaric acid by 22 times.

However, for many target molecules, especially those that are synthetic, no natural biosensor may exist. We developed a platform to engineer natural allosteric transcription factors with specificity to new inducer molecules. We computationally design for binding, synthesize and clone in multiplex thousands of specified sequences, and use a bidirectional screen to identify new responsive variants that retain allostery. We demonstrate by generating *E. coli* LacI variants responsive to gentiobiose, fucose, lactitol and sucralose. We uncovered significant plasticity in the ligand recognition of LacI, which may be a hallmark of allosteric transcription factors. Our method relies only on protein structure and operator DNA sequence, making it applicable to many other proteins.

These methods together advance the ability to engineer microbial biosynthesis of any target molecule using evolution. Additionally, designer transcription factors can enable broad applications from dynamic metabolic control to cell biology.

Table of contents

Acknowledgements

I owe thanks to a number of fine people and scientists for my professional success, and for enriching my life, during my time at Harvard Medical School.

First, my advisor, George Church. George keeps a youthful exuberance about doing big, field-shifting science, and it's infectious. He doesn't train students so much as he inspires them. As we Church lab members (eventually) leave and go on to other things, we form a diaspora of the ideals and maxims that George's style of science embodies. Build new tools. It is good to publish first, but better to publish last (i.e. solve the underlying problem). Enjoy the process (or in the words of Seth Shipman, "to keep spotting balloons.") In short: be fruitful and multiplex. (To multiplex means to turn what would normally be a single experiment into a collection of parallel variants of that experiment, together in one vessel, usually to save time, effort and cost simultaneously). In a way, George has created through his lab a multiplexing of his own powerful brand of scientific inquiry, which we each strive to emulate, while focusing through the lens of specific problems of interest to each of us. It has been an unparalleled, fun and enriching experience to spend this time in his lab, and I extend a very sincere thank you to George for letting me be a part of *his* grandest experiment.

George has attracted a group of wonderful and talented graduate students, postdocs, visitors, ethicists, artists and others too numerous to enumerate, but some of whom I must thank by name here.

Vatsan Raman, a postdoc with whom I collaborated on nearly all of my projects during graduate school. For his boundless data optimism even in the face of my boundless data

skepticism, for great discussions on protein evolution and politics, and for a bench I knew would likely be free to pour plates upon, I must thank him.

Jamie Rogers, a fellow Church graduate student, friend, and constant collaborator. For a wealth of ideas scientific and otherwise, great stories from abroad, for reminding me to take an engineer's approach to biology but to not spend my entire life in the lab, thank you.

Dan Goodman, whose massive undertakings of design, coding, sequencing, and bioinformatics analysis always made my projects feel pitiful and lackluster by comparison. For this dose of perspective, for years of good banter and discussion, and for even better scientific advice, I thank him.

I thank Alex Garruss for his energy, creativity, and optimism about all things biosensors and beyond. I thank both Alex Chavez and James DiCarlo, for their inspirational science, their advice, and their shared obsession with reef aquaria. I also thank Wei Leong Chew, Dan Mandell, Justin Feng, Raj Chari, Nikolai Eroshenko, Sri Kosuri, Marc Lajoie for comradery, scientific advice or softball pointers.

I extend a warm thank you to the multitude of unnamed Church lab members that have taught me so much about how little I know, how to approach hard problems and conquer them, or fail and attempt a new and better strategy, and who make the lab such a vibrant and peerless place. I thank my three great summer students for teaching me that I should not be trusted with any more summer students.

Thank you to my degree advisory committee members, Jagesh Shah, Greg Stephanopoulos and especially Mike Springer, who is also the chair of my dissertation

committee, for their guidance through my PhD process. I also kindly thank my readers and examiners, Kristala Prather, Neel Joshi and Simon Dove.

To my very good friends, in Boston and elsewhere, thank you for very much laughter, great connection regardless of how far away you live, and for keeping me sane either through your help, or by comparison.

To my BBS class, thank you for your sense of whimsy, for laughter, and for making me feel like a celebrity. To the BBS staff, especially Maria, Kate and Danny, thank you for bringing a very warm touch to a job you all do so well running the program.

I must thank my parents, Nancy and Lou, for giving me life. And I thank my whole family for love, support, and letting me run away to the other coast at 18 and stay to pursue my dreams. To my wife's family, my mishpucha, thank you for your love, support, and fantastic Whatsapp chatter.

Nearly last, but most importantly, I must thank my lovely wife, Adina. She forced me to learn how to communicate my work and why it is important to those far outside my grad school bubble, which gave me perspective and helped to shape my research. She has been a true partner in this process, a sounding board, a PowerPoint slide chastiser and fixer, and my support in every possible way. She has forgiven me for those countless times that "one more hour in the lab" certainly was not, and has still been more excited than me for every little success along the way. Thank you, darling.

Finally, I thank my daughter Lyla, for six months of constant, delightful distraction, and her childlike sense of wonder and curiosity at everything she sees. She reminds me that anything can be fascinating when seen through the right eyes.

Chapter 1. Background and context of the thesis

Attributions

Portions of this chapter are in preparation for submission as a review article to Current Opinion in Biotechnology, with Jameson K. Rogers and George M. Church.

Introduction

Enzymatic processes of cellular metabolism perform chemical conversions with exquisite specificity and speed. By engineering metabolism, these processes can be harnessed for human needs, such as the industrial production of organic chemicals, fuels, and polymers. Indeed, a large and rapidly growing segment of the 'bioeconomy' [1] comprises these microbial metabolic products. Despite this promise, metabolic engineering faces significant challenges to mature as an engineering discipline: biological 'parts', such as genes that encode enzymes, or promoters that direct their expression, can be highly context-dependent; and measuring the amount of a chemical produced requires chromatographic methods that are expensive and slow [2]. Because of this, design-build-test cycles are painstaking, and years to decades have so far been required to fully engineer the metabolism of an organism to produce a target chemical at commercially viable scale [3-5].

As a field, metabolic engineering has begun to shift toward an evolutionary paradigm that better suits the complexities of the problem. Traditional engineering disciplines have had great success in a bottom-up approach to design, in which well characterized parts follow well defined physical rules; by understanding the pieces, a whole system can be designed that behaves much as theory and calculation predict. Biological and metabolic engineering began

with a similar approach, and rules have been elucidated that govern how many essential parts behave, like ribosome binding sites [6], promoters [7], and even protein active sites [8], enabling forward design. However, efforts to design whole organisms using bottom-up approaches are rudimentary [9], and metabolic engineering must still rely on making modifications to organisms containing thousands of interacting genes, hundreds of which are of unknown function. However, biological systems are well suited to top-down design. If a clear objective can be set through a screen or selection, an optimal combination of genetic parts can be determined without comprehensive knowledge of each individual part.

For metabolic engineering, a cellular objective for efficient production of a target molecule can be set by deploying genetic devices, or biosensors, within cells to measure the target metabolite. These biosensors form a crucial interface that transmits information about the presence of a target molecule within a cell into actuation of genes that can report this information outside the cell for screening, or alternatively, express genes to change the behavior of that cell. By expressing fluorescence when a chemical is produced at high titer, for example, the biosensor allows bright, high-producing cells to be collected from a diverse population in a high-throughput manner. The single-cell resolution and high throughput identification afforded by a metabolite biosensor allows millions of variants to be assayed in multiplex in the same experiment. In this way, the engineering design-build-test cycle can be transformed into an evolutionary mutate-test cycle, with the engineer testing huge numbers of metabolic pathway designs in a rapid, iterative manner. This is the promise of biosensor-directed metabolic evolution, and this thesis presents significant steps toward the realization of allosteric transcription factors as biosensors to evolve metabolic pathways in *Escherichia coli*.

Metabolic design and modification strategies

To understand the need for multiplex evaluation of metabolite production within single cells, it is instructive to inspect the metabolic design and genomic modification capabilities that are now available, including their development, scale, and throughput. It is the advances in these 'design' and 'build' steps that require high throughput tools to test their success at engineering metabolism.

De novo pathway design

As the understanding of metabolic networks has matured [10], a number of techniques have been developed that can harness the immense wealth of information about the collective metabolism of the biosphere [11]. For target metabolic products not naturally produced by a host strain, algorithms can predict possible biosynthetic routes by borrowing enzymes from many organisms [12, 13]. These techniques are especially helpful where several missing enzymatic steps may be required. Such *de novo* biosynthetic pathways have enabled biological production of compounds including industrial feedstocks and pharmaceutical precursors like succinate [14], 3-hydroxyvalerate [15], glucaric acid [16], flavonoids [17-19], amorphadiene [20], and many others. A notable example was the engineering of *Escherichia coli* to produce the non-natural plastic precursor, 1-4-butanediol (BDO), requiring addition of five exogenous enzymes, and ultimately resulting in a strain capable of producing 18 g/L BDO after extensive engineering [4]. In this study, over 10,000 pathways were predicted, but these were narrowed down to just two candidates chosen for construction, using predicted yield, thermodynamics, and other constraints.

Strain optimization

Once a strain can produce a target chemical, as validated by liquid chromatography or mass spectrometry, an array of methods can be used to improve production titer, rate and yield. Titer is the concentration at which the target molecule accumulates, which can become toxic or inhibit enzyme activity; rate denotes that velocity at which reactants traverse through the pathway to form the product; and yield measures the proportion of pathway inputs that are converted into products, which is effected by stoichiometry, energetics, and thermodynamics [21]. The simplest pathway modifications include codon optimization of enzymes for higher expression or solubility in the new host, or overexpression using high plasmid copy numbers and strong promoters derived from phage.

These simple interventions may have limited success, and a wealth of more powerful methods exist to computationally inspect the entire host metabolic network [11]. OptKnock, one systems-level approach, looks for combinations of gene knockouts that may reduce byproduct formation, or alter cell metabolism in non-intuitive ways to maximize a target metabolite [22]. A related approach, flux-balance analysis (FBA)[23], models cellular metabolism as a system of linear equations of reactants and products to predict directional gene expression changes (up or down) that impact the production of a target product. From these early methods, dozens of variations have arisen [11, 24].

Careful choice of enzyme orthologues can also yield dramatic improvements in pathway performance, due to improved expression and solubility, better host cofactor compatibility, or better selectivity and kinetics. In the optimization of *E. coli* flavonoid biosynthesis, a tyrosine-ammonia lyase (TAL) from *Rhodotorula glutinis* showed more than 12-fold greater activity than

TAL from other sources [19]. When Bayer et al. [25] synthesized 89 methyl-halide transferase genes and tested their function in *E. coli*, they observed 56-85% having detectable activity on different halide ions, and saw several orders of magnitude range in activity. As DNA synthesis costs continue to decrease, building multi-step synthetic metagenomics pathways is becoming accessible, but will require very high-throughput evaluation to test each pathway combination.

Enzyme engineering is a daunting task because the space of multiple mutations increases rapidly with the number of residues that are addressed. An active site consisting of just five positions requires several million mutants to fully randomize, while saturation mutagenesis of seven positions requires over a billion mutants. Excitingly, computational design of new enzyme active sites has been demonstrated [8, 26, 27], but the new activities are often poor. Indeed, a computationally-designed retro-aldolase [26] showed several thousand fold improved activity when functional screening evolved a variant of the designed enzyme that exhibited drastic, unexpected active site remodeling [28].

A common theme among metabolic design methods is the ability to predict thousands or millions of high quality guesses toward which combinations of metabolic changes to an organism will yield the most productive strains, a great improvement over a literally infinite space of random mutations [29]. But these metabolic designs must be cloned and tested to identify productive variants, and to validate design methods for further improvement.

Genome engineering

Crucially, techniques to make mutations to microbial genes and genomes in targeted and multiplex fashion are becoming mature. Oligonucleotide-mediated genome editing in *E. coli*, called multiplex automated genome engineering (MAGE), makes use of the phage lambda

beta protein to integrate oligonucleotides bearing desired changes in the place of Okazaki fragments during DNA replication [30-32]. CRISPR-based genome editing techniques will help to expand genomic modification techniques to other bacteria [33] and yeast [34, 35]. Global transcription machinery engineering is a more random but holistic method to greatly diversify gene expression, which has demonstrated success in increasing *E. coli* tyrosine production [36], or improving *S. cerevisiae* ethanol tolerance and production [37] or resistance to fermentative stress [38]. Multiplex modification of genomes allows for the millions of combinations of gene expression changes generated by systems metabolic methods to be constructed *in vivo*.

Analogously, advances in DNA synthesis, especially microarray-synthesized oligonucleotide pools [39], and the assembly of these pools into full-length genes [40] are enabling the construction of rationally-designed gene libraries and collections of metagenomically-mined orthologues. Competition among companies supplying raw oligonucleotide pools – Agilent Technologies, Custom Array and Twist Biosciences – and others supplying fully synthesized genes – DNA 2.0, Genscript, IDT, Genewiz, Blue Heron, and more – is driving down the price of DNA [41] and transforming the capability of metabolic engineers to encode *in silico* designs into physical DNA.

The role of biosensors in metabolic engineering

Chemical measurement is a screening bottleneck

Without high-throughput or multiplex assays to identify productive strains, the power of computational predictions of new or optimized metabolic pathways, rationally-designed enzymes, and the ability to build billions of these high quality designs *in vitro* and *in vivo*, will

remain unrealized. The gold standard of metabolite measurement, using liquid chromatography or mass spectrometry, is limited to around 10^3 measurements per instrument per day with even the best custom setup [42], and more commonly, to fewer than 10^2 per instrument per day. One hundred thousand carefully constructed pathway variants, or one hundred million computationally predicted enzyme active sites are wasted if only one hundred of these can be assayed for function.

Conspicuous molecules, which are colorful or fluorescent, or aid cell fitness, illustrate the power of multiplex screens: Wang, et al. were able to optimize the bioproduction of lycopene, a bright red carotenoid, by generating an estimated 15 billion unique genetic variants and visually screening almost 10^6 of these to identify mutants with the highest reported production titer in just three days [30]. Most molecules of interest, though, lack such convenient spectroscopic detection, and are not essential for cell growth. For this majority, a mechanism is required to couple the presence of the inconspicuous molecule to a conspicuous reporter or fitness advantage.

Biosensors let cells make chemical measurements

A genetically-encoded biosensor can propagate molecular recognition of a target molecule into biological actuation within a cell, such as a change in protein activity or gene expression. In this way, each cell is enabled to 'measure' the amount of a target cell present, and can report this via a conspicuous gene product, providing the engineer with a multiplex method to detect cell biosynthetic productivity. Biosensors can be gauged for effectiveness via several metrics, with ideal biosensors showing appropriate operational concentration range, wide dynamic range, low false positive or 'escape' rate, and high molecular specificity [2].

Operational concentration is the range of target molecule concentrations over which the biosensor shows a graded, concentration-dependent change in response. The desired operational concentration varies by application: to detect new enzyme activity, nanomolar sensitivity might be ideal, but would be useless in discerning between cells that produce 0.1 versus 1.0 grams per liter of a target compound. In Chapter 2, we present an intervention to shift biosensor operational range by controlling export of the target chemical [43].

Dynamic range refers to the amplitude of the change in downstream signal, measured as the highest signal observed due to induction, minus the lowest un-induced baseline signal. Dynamic range can be affected by the number of copies of the biosensor within the cell [44], or by signal amplification using an enzymatic reporter [45]. The larger the dynamic range, the more reliably a true signal can be discerned from noise. In this way, dynamic range is related to the false positive rate, because a hundred-fold increase in signal is far less likely to be a noise artifact than a two-fold increase, at least within the same system.

False positives arise when spurious transcription, translation [46] or protein activity changes create a false signal that is not related to molecular detection. This dictates the number of designs that can be interrogated for rare winners. If one in 1,000 cells exhibits a false positive signal, searching for productive variants that exist at one in every 10,000 cells will yield a majority of detected cells that are not actually productive. Multiple cycles of enrichment may help, but when the biosensor output is a fitness advantage, such as antibiotic resistance, cells that escape selection may outgrow the population, amplifying their effect. In Chapter 2, we demonstrate a number of genetic biosensor modifications, as well as targeted counter-selection, to drastically decrease biosensor escape rate within a selection [43].

The most crucial biosensor characteristic is molecular specificity. For biosensor-directed metabolic engineering to be possible, a biosensor must be available for the target molecule that operates in the host organism of choice. Fortunately, cells have evolved a wide array of mechanisms to sense metabolite concentrations or cell states and respond to maintain homeostasis or dynamically alter their behavior.

Generation of new biosensors

Inducer molecule specificity is conferred by the binding affinity of the sensor domain chosen, and much effort has gone into identifying a wide array of sensory domains for sugars, amino acids, fatty acids and hydrocarbons, vitamins, and secondary metabolites including polyketides, macrolides, terpenes, flavonoids, and many others (Table 1.1).

Table 1.1 List of natural and engineered biosensors by molecule sensed. Abbreviated sensor type names refer to the following: allosteric TF, allosteric transcription factor; two-component, two-component systems; FRET, fluorescence resonance energy transfer; PBP, periplasmic binding protein; LBD, ligand-binding domain.

Table 1.1 (Continued)

Molecule(s)	Molecule type	Sensor	Sensor type	Reference
Natural biosensors				
1-butanol	Fatty alcohol, fuel	BmoR	Allosteric TF	[47]
Acrylic acid	Feedstock	AcuR	Allosteric TF	[44]
Adipate	Dicarboxylic acid	PcaR	Allosteric TF	[47]
Arabinose	Sugar	AraC	Allosteric TF	[45]
B12	Vitamin	BtuB	Riboswitch	[48]
Benzoic acid, naphthalene	Aromatics	NahR	Allosteric TF	[49]
Erythromycin	Macrolide	MphR	Allosteric TF	[50]
Fatty acids	Fatty acid	FadR	Allosteric TF	[51]
Galactose	Sugar	GalR	Allosteric TF	[52]
Glucaric acid	Feedstock	CdaR	Allosteric TF	[53]
Lactose	Sugar	LacI	Allosteric TF	[54]
Lysine	Amino acid	LysR	Allosteric TF	[55]
Muconate	Dicarboxylic acid	BenM	Allosteric TF	[56]
NADPH	Redox	SoxR	Allosteric TF	[57]
Naringenin	Flavonoid	TtgR	Allosteric TF	[58]
Octane	Alkane	AlkS	Allosteric TF	[59]
Ribose	Sugar	RbsR	Allosteric TF	[60]
Succinate	Dicarboxylic acid	DcuR	Two-component	[47]
Tetracylines	Polyketides	TetR	Allosteric TF	[45]
Xylose	Sugar	XylR	Allosteric TF	[61, 62]
Biosensor engineering examples				
Acyl-homoserine lactone specificity	Quorum sensing	LuxR	Allosteric TF	[63, 64]
Biphenyl, nitrotoluenes	Aromatics	XylR	Allosteric TF	[61]
Fucose, lactitol, gentiobiose, sucralose	Sugars, saccharides	LacI	Allosteric TF	See Chapter 3
Lactate	Sugar	Various	PBP	[65]
Mevalonate	Isoprenoid precursor	AraC	Allosteric TF	[66]
Progesterone, digoxin	Steroids	DIG10.3 LBD	Conditional stability	[67]
Pyruvate	Alpha-keto acid	*De novo*	FRET	[68]
Serotonin	Neurotransmitter	Various	PBP	[65]
SLF*-derivatives	Synthetic polyaromatics	FKBP12	Conditional stability	[69]

Table 1.1 (Continued)

Molecule(s)	**Molecule type**	**Sensor**	**Sensor type**	**Reference**
Theophylline	Alkaloid	*De novo*	Riboswitch	[70]
Thiamine pyrophosphate	Vitamin	*De novo*	Riboswitch	[71, 72]
Trehalose-6-p	Sugar	*De novo*	FRET	[73]
Triacetic acid lactones	Feedstock	AraC	Allosteric TF	[74]
Trinitrotoluene	Aromatic	Various	PBP	[65]
Vanillin	Aromatic	QacR	Allosteric TF	[75]
Zn^{2+}	Ion	*De novo*	FRET	[76]
3,4-dihydroxybenzoate	Aromatic	PobR	Allosteric TF	[77]

Previously developed small molecule biosensors include many mechanistic paradigms, and are overwhelmingly co-opted from the natural sensory machinery of cells (Table 1.1). Allosterically-regulated transcription factors (aTFs) change their affinity for an operator DNA sequence through a conformational change enacted by ligand binding, and are useful to directly control gene transcription [45, 52, 78, 79]. Ligand-dependent protein dimerization [80] and ligand-conditional protein stability [69] are related methods that require binding of a small molecule to stabilize a protein dimer interface, or a protein monomer to avoid degradation, respectively; these methods can lead to direct changes in fluorescent reporter protein function, or mediate transcriptional changes via an additional two-hybrid system [81]. Riboswitches are 5' untranslated regions composed of RNA that binds to a small molecule ligand and controls the stability of mRNA transcripts, directly affecting the translation of the encoded genes [82-86]. Fluorescence resonance energy transfer (FRET)-based sensors [76, 87] use a conformational change in a protein domain brought about by ligand binding to change the proximity of two fluorophores capable of excitation-emission photon transfer for direct ligand detection. Two-component systems [88], though ubiquitous in bacteria, are of limited use for intracellular biosensors, as they chiefly mediate signal transduction of extracellular ligand detection, but may be useful for quorum sensing [89] or other bulk properties of a culture or fermentation. The function of new sensory gene family members can often be inferred in bacteria due to their proximity to the operons they regulate, and metagenomic sequence mining will continue to expand the repertoire of natural sensors available [90, 91].

Where natural sensory domains have not been found, or likely have yet to evolve (i.e. a synthetic target molecule), sensor engineering strategies can expand or alter the molecular

specificity of existing sensors, or even create new ones *de novo* (Table 1.1). Periplasmic binding proteins (PBPs) have been altered to bind new substrates and report binding through a fluorescent signal created by dye-labeled cysteines [65]. Random mutagenesis, or saturation mutagenesis of key positions, has shown promise in changing specificity of allosteric TFs, including AraC to sense mevalonate [66], or LuxR to change specificity [63, 64]. Computational design methods can sample a much richer set of mutations than can random approaches, and computational design of ligand binding interfaces has shown success [92]. By incorporating designed binding interfaces into a doman requiring target molecule binding for protein stability, sensors for new molecules can be created, including progesterone and digoxin [67]. For allosteric TFs, structure-based computational design incorporating homology modeling [77] or mechanistic insights [75] has enabled the development of new sensors for 3,4-dihydroxybenzoate or vanillin, respectively. In Chapter 3, we present a general strategy to redesign the ligand specificity of allosteric TFs using computational design and bidirectional screening [93].

Biosensor implementation

For metabolic engineering, biosensors are generally employed in one of two ways: as sensor-reporters or sensor-actuators. A sensor-reporter transfers information about the intracellular presence of a chemical outside the cell via expression of reporter genes with a detectable phenotype [49]. In contrast, sensor-actuator dynamically modifies the metabolism of a cell based on the presence of a target chemical, pathway intermediate or cell state [94].

Sensor-reporters for multiplexed phenotype evaluation

Sensor-reporter screens have been demonstrated using a number of different reporters: fluorescence, insoluble pigments, luminescence, and antibiotic resistance. In one of the earliest examples, van Sint Fiet, et al. [49] used NahR, a transcriptional activator induced by the benzoic acid product formed by a benzaldehyde dehydrogenase, to express LacZ; this enabled blue X-gal products to report on benzoic acid production, allowing detection of productive cells at a frequency of 10^{-6}. Regulating fluorescent protein expression, sensor-reporters have been used to screen for increased microbial production of the isoprenoid precursor mevalonate [66], L-lysine [95], 1-butanol [47], and triacetic acid lactones (Tang, 2013). Sensor-coupled antibiotic selections use cell fitness as a proxy for target metabolite production, which has been used to identify improved 1-butanol production plasmids [47], or even for whole-pathway iterated selection to evolve higher production of glucaric acid or naringenin [43] (see Chapter 2). Selection enables large library sizes theoretically limited only by the size of the culture, but the rate of selection escape often imposes practical restraints that are far lower, and must be addressed [43] (see Chapter 2). Luciferase has been used as a reporter to screen for production of macrolides [96], or to detect toluene and related compounds [97]. Together, these works demonstrate that sensor-reporters are a viable strategy for screening to improve metabolic pathways.

Sensor-actuators for dynamic metabolic control

Beyond simply reporting the presence of a target chemical, biosensors also have been integrated in a more physiological context to dynamically modify host cell metabolism [98].

Noting that deregulated or uncontrolled pathways are often created through gene knockouts, Farmer and Liao [94] repurposed one of the global regulatory systems in *Escherichia coli*, the Ntr regulon, to control an engineered lycopene biosynthesis pathway, expressing key genes in response to the presence of acetyl phosphate, which is a signal of excess flux through glycolysis; this intervention significantly increased lycopene biosynthesis. In a powerful demonstration of dynamic pathway regulation, production of fatty acids and derived metabolic products was boosted threefold to 28% of theoretical yield, by using the *E. coli* endogenous regulator, FadR, to express two key enzymes in response to the presence of excess acyl-CoA [51]. Tapping into central, essential metabolic functions of cells, means that interventions can starve the cell of resources needed to build biomass. By regulating enzymes to produce biodiesel and other fuel molecules in response to malonyl-CoA pools, this essential intermediate could be safely siphoned into fuel production, boosting yields without harming cell viability [99, 100]. A similar approach was successfully implemented in yeast, by expressing squalene synthase in response to glucose availability, boosting sesquiterpene production [101].

Many metabolic pathways producing non-native molecules, or native molecules at high titer, lead to toxic intermediates that retard cell growth. Dynamic biosensor-actuators can effectively alleviate this toxicity, if sensory domains responsive to the toxic product can be identified [102]. Of great utility in dense fermentation cultures, dynamic regulation of biofilm dispersal, or expression of toxic production enzymes once maximal biomass has been reached, can be controlled by harnessing natural quorum sensing mechanisms present in many bacteria [103].

The future of biosensor-directed metabolic engineering

Biosensors offer an attractive, multiplexed phenotype screening solution with the potential to revolutionize metabolic engineering. Thus far, bionsensor-mediated pathway production gains have been modest, and have not approached the grams per liter production titers typical of mature pathways required for commercial scale (e.g. [4]). As an emerging field, most studies have been proof of concept in nature, targeting a small number of genes, and using a single round of screening with a single biosensor. The strategies are clever and promising, but not yet ready for industrial use.

To mature as a field, biosensor-directed metabolic engineering needs an integration of dynamic metabolic control and screening approaches, for example using several biosensors for pathway intermediates that produce key enzymes only when needed, while allowing a final product-specific sensor-reporter to allow screening of top pathway designs. Full biosynthetic pathways must be targeted [43, 95], and multiple sensors may be required, with graduated operational concentration ranges to avoid saturating the biosensor at high molecule titers. To enable very large libraries, of 10^9 members and above, new interventions to improve robustness to false positives will be required, which may benefit from standardized screening chassis [44]. Following these recommendations, metabolic engineering will benefit from a powerful application of evolutionary strategies that are ideal to solve this difficult class of biological problems.

Chapter 2. Evolution-guided optimization of biosynthetic pathways

Attributions

The majority of this chapter was previously published in the Proceedings of the National Academy of Sciences [43] and does not require explicit permission to reproduce here.

Srivatsan Raman, Jameson K. Rogers, Noah D. Taylor and George M. Church conceived the project, analyzed the data, and reviewed the manuscript. Srivatsan Raman, Jameson K. Rogers and Noah D. Taylor designed and performed all experiments, wrote the manuscript, and contributed equally to this work.

Summary

Engineering biosynthetic pathways for chemical production requires extensive optimization of the host cellular metabolic machinery. Because it is challenging to specify *a priori* an optimal design, metabolic engineers often need to construct and evaluate a large number of variants of the pathway. We report a general strategy that combines targeted genome-wide mutagenesis to generate pathway variants with evolution to enrich for rare high producers. We convert the intracellular presence of the target chemical into a fitness advantage for the cell by using a sensor domain responsive to the chemical to control a reporter gene necessary for survival under selective conditions. Because artificial selection tends to amplify unproductive cheaters, we devised a negative selection scheme to eliminate cheaters while preserving library diversity. This scheme allows us to perform multiple rounds of evolution (addressing ~10^9 cells/round) with minimal carryover of cheaters after each round. Based on candidate genes identified by flux balance analysis, we used targeted genome-wide

mutagenesis to vary the expression of pathway genes involved in the production of naringenin and glucaric acid. Through up to four rounds of evolution, we increased production of naringenin and glucaric acid by 36- and 22-fold, respectively. Naringenin production (61 mg/L) from glucose was more than double the previous highest titer reported. Whole genome sequencing of evolved strains revealed additional untargeted mutations that likely benefit production, suggesting new routes for optimization.

Introduction

Microbial production of chemicals presents an alternative to ubiquitous chemical synthesis methods. Biosynthetic production is attractive because it can utilize a broad assortment of organic feedstocks, proceed under benign physiological conditions, and reduce environmentally deleterious byproducts. Biosynthetic alternatives are being pursued for a wide range of chemicals, from bulk commodity building blocks to specialty chemicals.

Natural cells are seldom optimized to produce a desired molecule. To achieve economically viable production, extensive modifications to host cell metabolism are often required to improve metabolite titer, production rate and yield. The optimizations of biosynthetic pathways for 1,3-propanediol [3], flavonoids [18, 19], L-tyrosine [104], and 1,4-butanediol [4] illustrate this complexity. Fortunately, computational models of cellular metabolism, such as flux-balance analysis (FBA), aid in predicting metabolic changes likely to improve the production of a target molecule. Powerful methods including oligonucleotide-directed genome engineering (MAGE) [30] and Cas9-mediated editing [33] can specifically mutate the genomic targets predicted by FBA. But the combinatorial space of these genomic

mutations quickly outstrips the throughput of current analytical methods for evaluating chemical production in individual clones (<10^3 samples machine^{-1} day^{-1}).

Biosensors that report on the concentration of a chemical within each individual cell can alleviate this screening bottleneck. Such sensor-reporters transduce the binding of a target small molecule by a sensory protein or RNA into a gene expression readout [49]. The resulting expression of a fluorescent reporter gene or antibiotic resistance gene allows facile identification of mutant cells with increased production of the target chemical.

Sensor-reporters have been employed to screen for increased microbial production of several chemicals, including the isoprenoid precursor mevalonate [66], L-lysine [95, 105], 1-butanol [47] and triacetic acid lactone [74]. These studies evaluated a set of variants that altered the expression or coding sequences of one or two key enzyme genes encoded on a plasmid [47, 66, 74, 105]. Similarly, a lysine-responsive sensor-reporter was used to uncover new endogenous enzyme mutants in *Corynebacterium glutamicum* implicated in higher L-lysine production [95].

We sought to expand the scope of sensor-directed metabolic engineering to the directed evolution of whole endogenous pathways. Using FBA as a guide, we simultaneously targeted up to 18 *E. coli* genomic loci to induce mutations in regulatory or coding sequence of genes implicated in biosynthesis of a target molecule. We established a robust selection, utilizing a sensor protein responsive to the target chemical to regulate the expression of an antibiotic resistance gene. Nearly a billion pathway variants could be evaluated simultaneously, enriching for the best producers when selection pressure was applied.

A major challenge faced by this selection approach (and a difficulty for most genetic selections) is the incidence of cheater cells that survive without producing the target molecule. These cheaters evolve to survive selection by mutating the sensor or selection machinery, rather than through higher target molecule synthesis. Lacking a metabolic burden, these 'evolutionary escapees' outcompete the top producers during a selection. Multiple selection cycles compound escape, obscuring productive cells and making further pathway evolution infeasible. We therefore devised a selection scheme that, by toggling between negative and positive selection, allows us to remove escapees from the population when they arise. This strategy maintained high selection fidelity, permitting multiple rounds of evolution to progressively enrich for higher producing cells.

For sensor-reporter metabolic engineering to be generalizable, sensor domains specific to many different target molecules must be available. Fortunately, natural sensors exist for a wide array of industrially-relevant chemicals, including aliphatic hydrocarbons, short-chain alcohols, sugars, amino acids, polymer building blocks, and vitamins. Many more sensor domains are likely to be present among the thousands of additional bacterial regulators known from sequence [106-108] that remain to be characterized. We adapted 10 regulators to our selection system, creating synthetic dependence on their cognate inducer molecules, and demonstrated the utility of two of these for genome-wide metabolic engineering.

Results

Sensor-selectors are a specific example of the sensor-reporter paradigm that use a gene whose product confers a fitness advantage (e.g. antibiotic resistance) as the reporter. Our

sensor-selector architecture encodes a chemical-responsive sensor domain together with its cognate promoter, which controls a selectable reporter (Figure 2.1A). We show that this general implementation is suitable for transcriptional regulators (both activators and repressors) and riboswitches, that collectively respond to a wide variety of chemicals (Figure 2.2A; Supplementary Table S2.1).

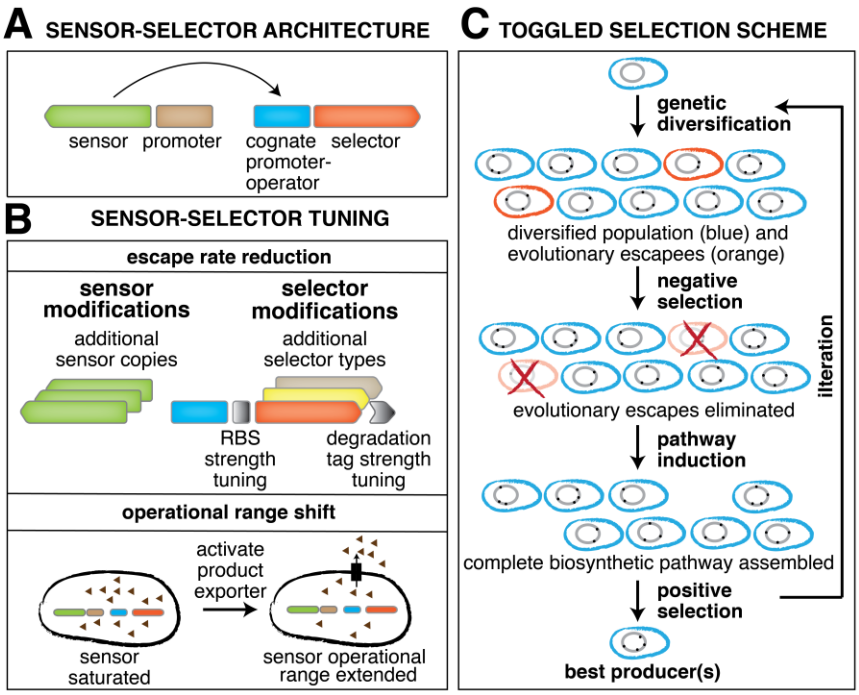
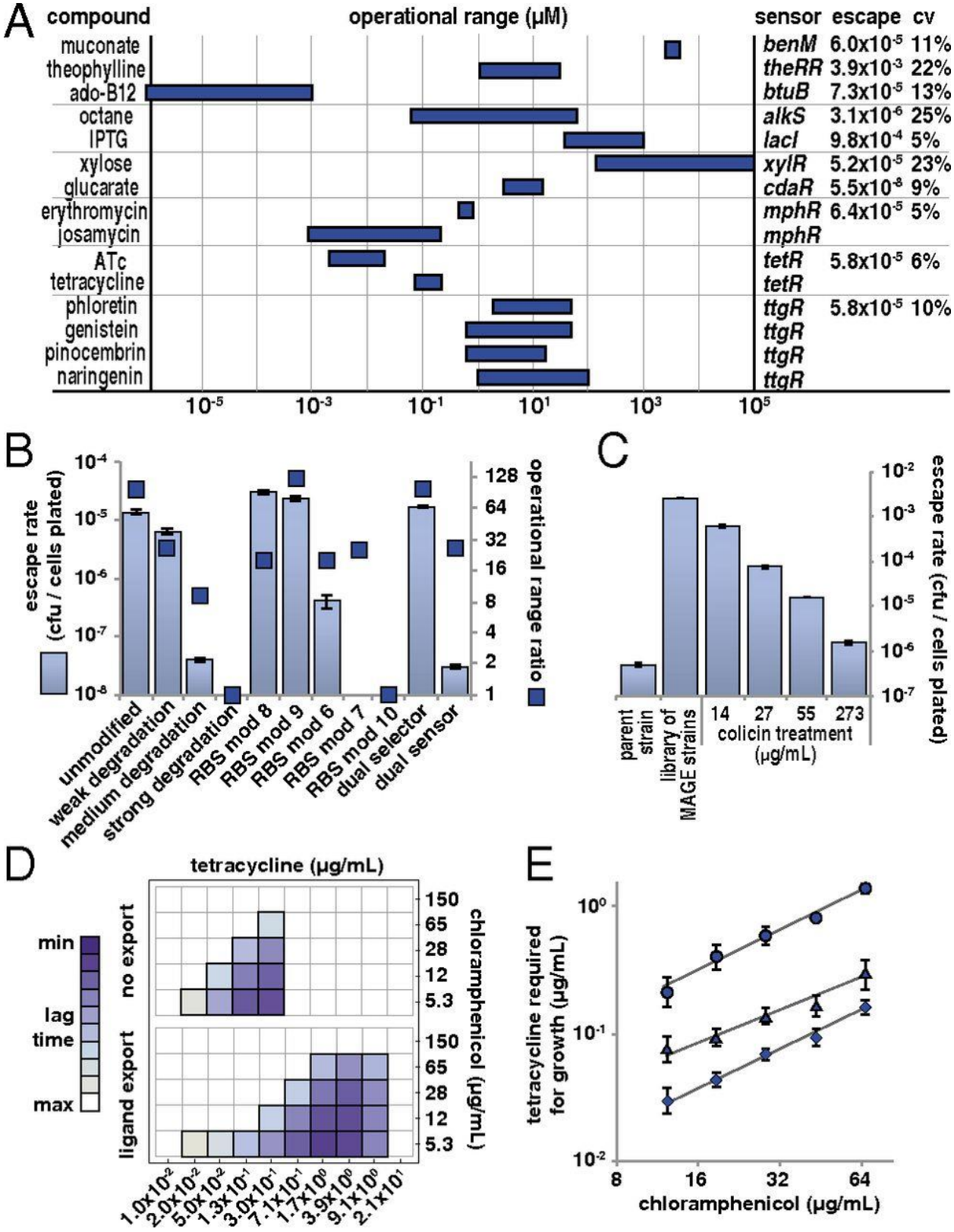


Figure 2.1 Sensor-selector design and pathway optimization through toggled selection (A) Sensor-selector genetic architecture. (B) Methods for tuning sensor-selectors to reduce escape rate and shift operational range. Escape rate is reduced by (1) adding a degradation tag, (2) mutating the RBS of the selector, (3) including multiple orthogonal selectors or (4) including an additional copy of the sensor. Activating an exporter shifts the sensor-selector operational range. (C) Toggled selection protocol for biosynthetic pathway optimization through multiple rounds of evolution. Negative selection eliminates cheaters; subsequent positive selection identifies higher producing clones from a diverse library.

Figure 2.2 Characterization of sensor-selector modifications. (A) Escape rate and operational range of ten sensors with cognate inducer chemicals and TolC as a selector. Horizontal bars depict the operational range. The lower bound of the range reflects the lowest concentration of exogenously supplied inducer that provides a selective advantage. The upper bound of the range indicates that higher inducer concentration does not increase fitness advantage. (B) Effect of genetic modifications on the TtgR-TolC sensor-selector escape rate and operational range. Escape rate (light blue bars, left axis) is the proportion of cells that evade selection (cfu/cells plated). Escape rate not shown if below the limit of detection (10^{-10} cfu/cells plated). Escape rate operational range ratio (blue boxes, right axis) is the ratio of the high concentration of the operational range to the low concentration of the operational range. (C) MAGE mutagenesis increases the escape rate (CFU/cells plated) in the CdaR-TolC strain. Treatment with colicin E1 removes escapees in a dose-dependent manner. (D) Tetracycline exporter (tetA) expression shifts the operational range of the TetR-CAT (chloramphenicol acetyltransferase) sensor-selector. Growth lag times reported for orthogonal concentration gradients of tetracycline vs. chloramphenicol in the absence of tetA (top panel) compared to tetA expression (bottom panel). (E) The shift in TetR-CAT operational range is tunable by titration of tetA expression. The minimum tetracycline concentration required for growth (y-axis) at a given selection pressure (x-axis) for three tetA expression levels: none (diamonds), intermediate (triangles), high (circles). Error bars represent S.E.M. of production from 3 biological replicates.

Figure 2.2 (Continued)



Each sensor-selector exhibits unique behavior, dependent on sensor affinity for the chemical, sensor type and induction response; for example, the escape rate and operational range can vary over orders of magnitude for different sensors (Figure 2.2A). For each sensor, the operational range is defined as the chemical concentration range over which cells continue to experience a marginal fitness advantage with increasing concentration. The lower bound of the range reflects the lowest concentration of exogenously supplied inducer that provides a selective advantage. The upper bound of the range indicates that higher inducer concentration provides no additional fitness advantage. This range informs the utility of a sensor for optimizing a pathway. We measured the operational range of ten sensor-selectors; the MphR [50], TtgR [58], and TetR [45] operational ranges were measured for multiple inducers (Figure 2.2A).

Under selection pressure, most cells in a sensor-selector strain population survive only when the target chemical is detected. But a small fraction of the cells survive absent the chemical. 'Evolutionary' escape results from mutations that permanently reduce selection sensitivity, and additionally, natural sensors may not have evolved to completely repress the basal expression level of the genes they regulate. In our selections, the resultant constitutive or leaky selector expression generates false positives, making it difficult to identify rare winners. Promoter engineering to optimize the placement of operator sites can yield very tight repression [45], but this approach requires specific development for each sensor. Instead, our standardized construction allows us to reduce the effect of leaky selector expression through common *cis*-regulatory modifications that are sensor-independent. These modifications include

appending a degradation tag to the selector to accelerate its proteolysis and mutating the ribosome binding site (RBS) of the selector gene to attenuate translation (Figure 2.1B).

We implemented several modifications in the TtgR-TolC sensor-selector strain for comparison. Appending *ssrA* degradation tag variants to TolC reduced escape, in correlation to the strength of the degradation tag [109], by as much as six orders of magnitude (Figure 2.2B). However, reduced escape also decreased the operational range. We adjusted the spacing between the RBS and translation start site of TolC to achieve fine-grained translation control [110]. Five of ten spacing mutations reduced escape rate while maintaining a measurable operational range (Figure 2.2B, Supplementary Figure S2.6). For a dual selector strain, in which TtgR regulates both *tolC* and a kanamycin resistance gene, observed escape rates support the hypothesis of escape through leaky reporter expression: with both SDS and kanamycin present, the escape rate was much lower ($5.2 \pm 0.21 \times 10^{-8}$ cell/cell) than with either SDS alone ($1.7 \pm 0.092 \times 10^{-5}$ cell/cell) or kanamycin alone ($4.4 \pm 0.44 \times 10^{-4}$ cell/cell). Finally, we observed substantial escape rate reduction using two copies of the *ttgR* sensor gene and a single TolC selector (Figure 2.2B). Because TtgR acts as a transcriptional repressor, evolutionary escape requires inactivating mutations to both gene copies, and higher sensor expression may reduce escape through tighter basal repression of the selector.

Sensors are useful for pathway optimization only when the intracellular concentration of the target chemical is within the operational range of the sensor. We hypothesized that expressing an exporter of the target chemical should decrease the intracellular concentration, shifting the operational range (Figure 2.1B). We studied this export effect by expressing a tetracycline exporter, TetA, in cells that place the tetracycline-responsive sensor, TetR in

control of chloramphenicol acyltransferase (CAT) expression. When this strain expressed *tetA*, the entire operational range for tetracycline, including both the lower detection threshold and upper saturation point, shifted about ten-fold higher (Figure 2.2D). This effect was tunable by controlling *tetA* expression from the arabinose-inducible pBAD promoter (Figure 2.2E). The CAT selector was used here due to improved titration of drug sensitivity.

Pathway evolution by toggled selection

To maximize the likelihood of identifying rare cells with a higher production phenotype, we developed a toggled selection scheme (Figure 2.1C) that preserves library complexity while eliminating evolutionary escapees. Evolutionary escapees are cells that acquire mutations to survive selection without producing the target chemical. This escape prevents the identification of rare winners in a selection, and confounds multiple rounds of evolution as these escapees outcompete the productive cells. Through toggled selection we can selectively kill the escapees at each round, and carry over the productive cells for further improvements in subsequent rounds. We chose TolC [111] as a selector because of its utility for both positive selection (using sodium dodecyl-sulfate; SDS) and negative selection (using colicin E1). MAGE is highly mutagenic, increasing the escape rate from below 10^{-7} to above 10^{-3} after five cycles in the CdaR-TolC sensor selector strain. This increase could be reversed by incubation with colicin E1 (Figure 2.2C), because evolutionary escapees evade SDS toxicity through mutations that constitutively express *tolC*, making them highly susceptible to colicin E1. Crucially, we ensure that productive cells are not also killed during negative selection by maintaining a pathway gene under tight transcriptional control, which prevents prematurely triggering the sensor (Figure 2.1C). After negative selection, we induce the regulated enzyme, allowing cells to

produce the target chemical, and the sensor expresses *tolC* in proportion to chemical production. By toggling to positive TolC selection with SDS, we enrich for higher producers, and these can be characterized for their production phenotypes or subjected to further pathway evolution (Figure 2.1C).

Naringenin pathway

We implemented the toggled selection scheme to evolve *E. coli* toward higher production of two chemicals: naringenin and glucaric acid. Naringenin, a pharmacologically useful plant flavonoid molecule, was chosen because previous efforts serve to benchmark our optimization [18, 19, 112]. *E. coli* requires four heterologous enzymes to synthesize naringenin from glucose: tyrosine ammonia lyase (TAL), 4-coumaroyl ligase (4CL), chalcone synthase (CHS) and chalcone isomerase (CHI) [19] (Figure 2.3A). Because this pathway consumes tyrosine and malonyl-CoA, our strain engineering strategy targeted endogenous *E. coli* gene regulatory and coding loci to increase the availability of these precursors (Supplementary Table S2.4). The scope of this work was genomic mutagenesis, so the heterologous genes were left untargeted.

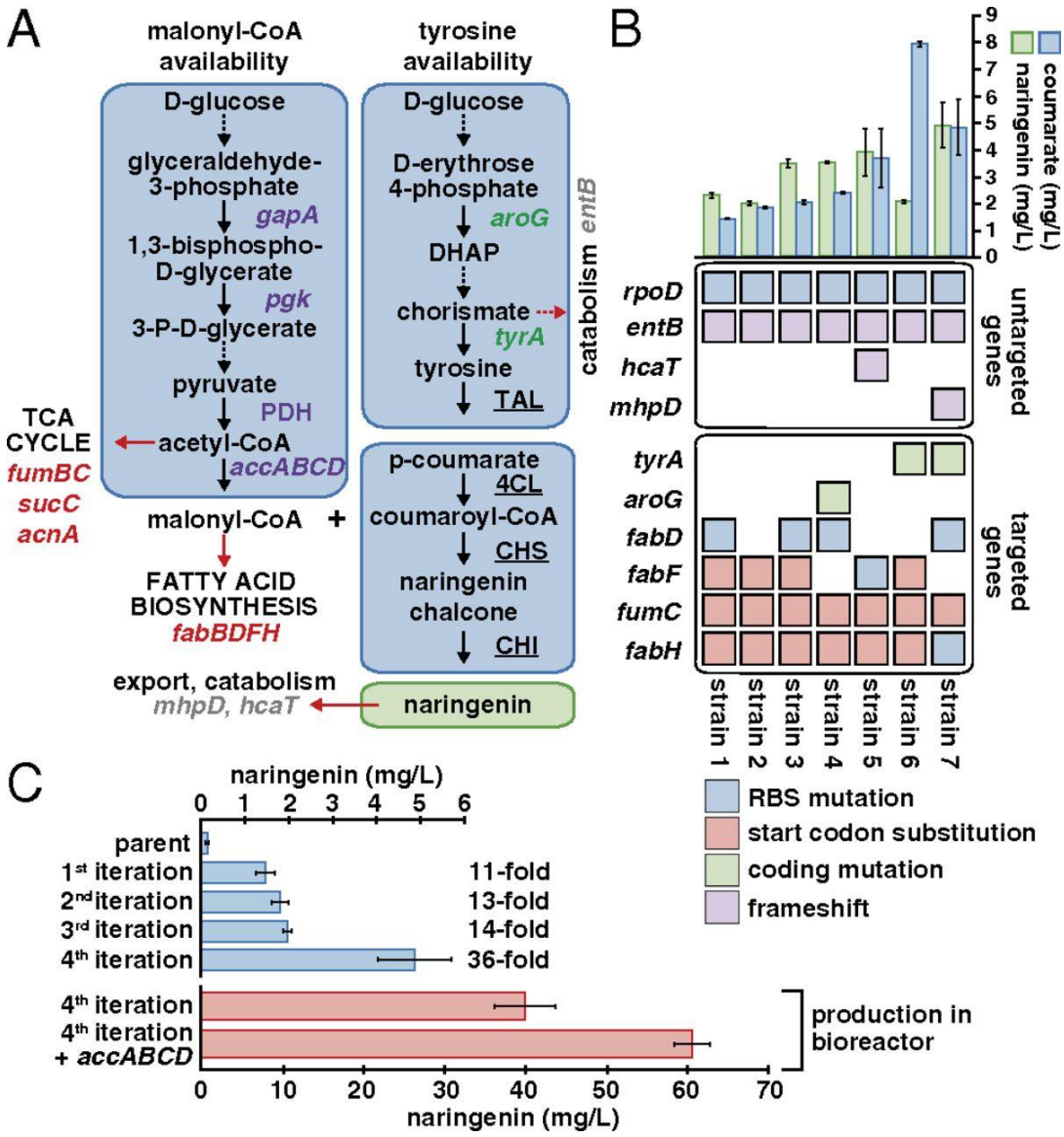


Figure 2.3 Optimization of the naringenin biosynthetic pathway. (A) Endogenous E. coli genes targeted by MAGE to increase malonyl-CoA and tyrosine availability for naringenin production; targeted genes are colored: purple, up-regulation; red, down-regulation; green, coding changes; gray, untargeted knocked out genes. (B) Genotype and production phenotype of the top seven producers (in no particular order) from the fourth round of toggled selection. Colored boxes denote the type of genetic modification. Shown are mutations found at targeted genes (Bottom) and those at untargeted genes (Center). Naringenin (green bars) and coumaric acid (blue bars) concentrations for single production measurements are shown above the corresponding genotype (Top). Error bars represent SEM of three biological replicates. (C) Average naringenin production titers for parent and highest producer after each round of evolution (blue bars). Production titer from fed batch bioreactor fermentation of the highest producer and highest producer with accABCD overexpressed (red bars).

We performed flux balance analysis (FBA) toward increased malonyl-CoA, because its availability limits naringenin production [18, 112] (Supplementary Table S2.6). FBA identified three key pathways: glycolysis, fatty acid biosynthesis and the tricarboxylic acid (TCA) cycle (Figure 2.3A; Supplementary Table S2.3). Greater flux through glycolysis by up-regulation of *gapA*, *pgk* and *pdh* should increase pools of acetyl-CoA, which is converted to malonyl-CoA by acetyl-CoA carboxylase enzymes *accABCD*. Because acetyl-CoA is oxidized in the TCA cycle, we targeted for down-regulation TCA enzymes *mdh*, *fumBC* and *acnAB*. To throttle acetyl- and malonyl-CoA consumption in fatty acid biosynthesis, we targeted *fabBDFH* for down-regulation. Availability of tyrosine, the other precursor for naringenin production, is limited by activity of two enzymes in aromatic biosynthesis, *aroG* [113] and *tyrA* [104] that are inhibited by 3-deoxy-D-arabinoheptulosonate 7-phosphate (DAHP) and chorismate, respectively. We targeted *aroG* and *tyrA* for coding sequence changes shown to alleviate product inhibition. These predictions (Figure 2.3A) corroborate interventions experimentally shown to increase production of malonyl-CoA [112], tyrosine [36] and naringenin [18, 19].

Previous efforts to engineer the naringenin pathway have relied on plasmid-based over-expression or complete knockouts [112]; for tightly-regulated or essential central metabolism genes, such drastic modifications can have deleterious growth defects. For finer control of gene expression states, which can more closely balance biosynthetic and survival objectives, we used multiplex automated genome engineering (MAGE) [30]. Oligonucleotides for MAGE mutagenesis were targeted to Shine-Dalgarno sequences to finely increase or decrease translation efficiency, to alternative start codons (CTG, GTG or TTG) to yield larger translational attenuation, or to introduce premature stop codons or coding frameshifts for complete

inactivation (Supplementary Table S2.4). Seven genes were identified by FBA for overexpression to increase flux through glycolysis and to convert acetyl-CoA to malonyl-CoA.

Four rounds of evolution by toggled selection were performed on the strain containing two copies of the *ttgR* sensor gene controlling TolC, due to its favorable combination of escape rate and operational range (Figure 2.2B). We verified that TtgR responds only to naringenin but not the pathway intermediate *p*-coumaric acid (Supplementary Figure S2.3). After four rounds, each consisting of about 15 cycles of targeted mutagenesis followed by toggled selection, the best strain identified produced 36 times more naringenin than the parent strain (Figure 2.3B). We screened approximately 20 colonies to identify the highest producer at each round. With a supernatant concentration of 39 mg/L, the production titer of this strain surpasses the highest published production of naringenin (29 mg/L) directly from glucose [19] (Figure 2.3B). We further enhanced the production titer to 61 mg/L by overexpressing *E. coli* acetyl-CoA carboxylase genes (*accABCD*), which have been shown to increase endogenous malonyl-CoA levels (Figure 2.3B; Supplementary Figure S2.1). Through genetic changes alone, we were able to nearly recapitulate the high naringenin titer (84 mg/L) previously achieved by addition of cerulenin [19], an inhibitor of fatty acid biosynthesis, which is prohibitively expensive for industrial scale production.

We sequenced the genomes of the starting strain and seven high-producing strains isolated after evolution round four. All seven strains incorporated RBS or start codon changes at several targeted loci (Figure 2.3C). We found a number of mutations associated with malonyl-CoA production (Figure 2.3C; Supplementary Table S2.7). In the TCA cycle, fumarase was down-regulated by a *fumC* start codon mutation in all seven strains (likely due to its selection in an

early round). Several fatty acid genes were also down-regulated. Fatty acid biosynthesis genes whose products initiate synthesis from acetyl-CoA (*fabH*) or malonyl-CoA (*fabD*) were down-regulated by start codon or RBS mutations in seven and four strains, respectively. The fatty acid elongation gene *fabF* had start codon attenuation (GTG to TTG) in four strains and a purine to pyrimidine mutation in the RBS predicted to lower translation rate [6] in a fifth strain (Figure 2.3C; Supplementary Table S2.7). None of the seven strains had a down-regulation target knocked out, and none of the strains had mutations affecting *fabB*, an essential gene, reflecting a balance between production and growth objectives. Computational prediction of translation rate shows that selected clones enrich for RBS and start codon mutations that attenuate translation of genes, consistent with FBA predictions (Supplementary Figure S2.4).

Three strains exhibited targeted mutations in tyrosine biosynthetic genes shown to alleviate product inhibition. All three produced substantially more coumaric acid, including two strains with the *tyrA* mutation A354V, which produced at least an order of magnitude more coumaric acid (Figure 2.3C). This large coumaric acid buildup suggests that malonyl-CoA may be limiting for naringenin production in these strains. In support of this idea, overexpression of the enzymes AccABCD, which convert acetyl-CoA to malonyl-CoA, increased naringenin production almost 1.5-fold in the evolved strain (Figure 2.3B).

While the MAGE process concentrates diversity generation on targeted loci and increases the probability of sampling specific mutations hypothesized to confer beneficial phenotypes, it also has unintended mutagenic effects. Whole genome sequencing revealed many non-targeted mutations in the producer strains (Figure 2.3C; Supplementary Tables S2.7 and S2.8), including several mutations likely involved in higher naringenin production.

Frameshifts inactivated *mhpD*, which catabolizes aromatic compounds similar to coumaric acid [114], and *hcaT,* a putative transporter of phenylpropionates like coumaric acid [115]. Similarly, a frameshift in *entB*, which diverts chorismate from aromatic biosynthesis, may increase tyrosine production [116]. We speculate that knocking out all three enzymes facilitates production of naringenin by increasing the concentration of the precursor, p-coumaric acid. Attributing function to non-coding regulatory mutations is more tenuous. However, we observed a mutation in the Shine-Dalgarno sequence of *rpoD*, mutation of which increases tyrosine production [36].

Glucaric acid pathway

In order to validate directed evolution by sensor-selectors as a generalizable method, we optimized the production of glucaric acid in *E. coli*. Glucaric acid was chosen for two reasons. First, unlike naringenin production, previous work to modulate endogenous pathways was absent. Second, glucaric acid was identified as a key renewable chemical for the replacement of petroleum-based polymer production. Glucaric acid can be synthesized in *E. coli* by expression of three exogenous enzymes: myo-inositol-1-phosphate synthase (Ino1), myo-inositol oxygenase (MIOX) and uronate dehydrogenase (Udh) [16] (Figure 2.4A).

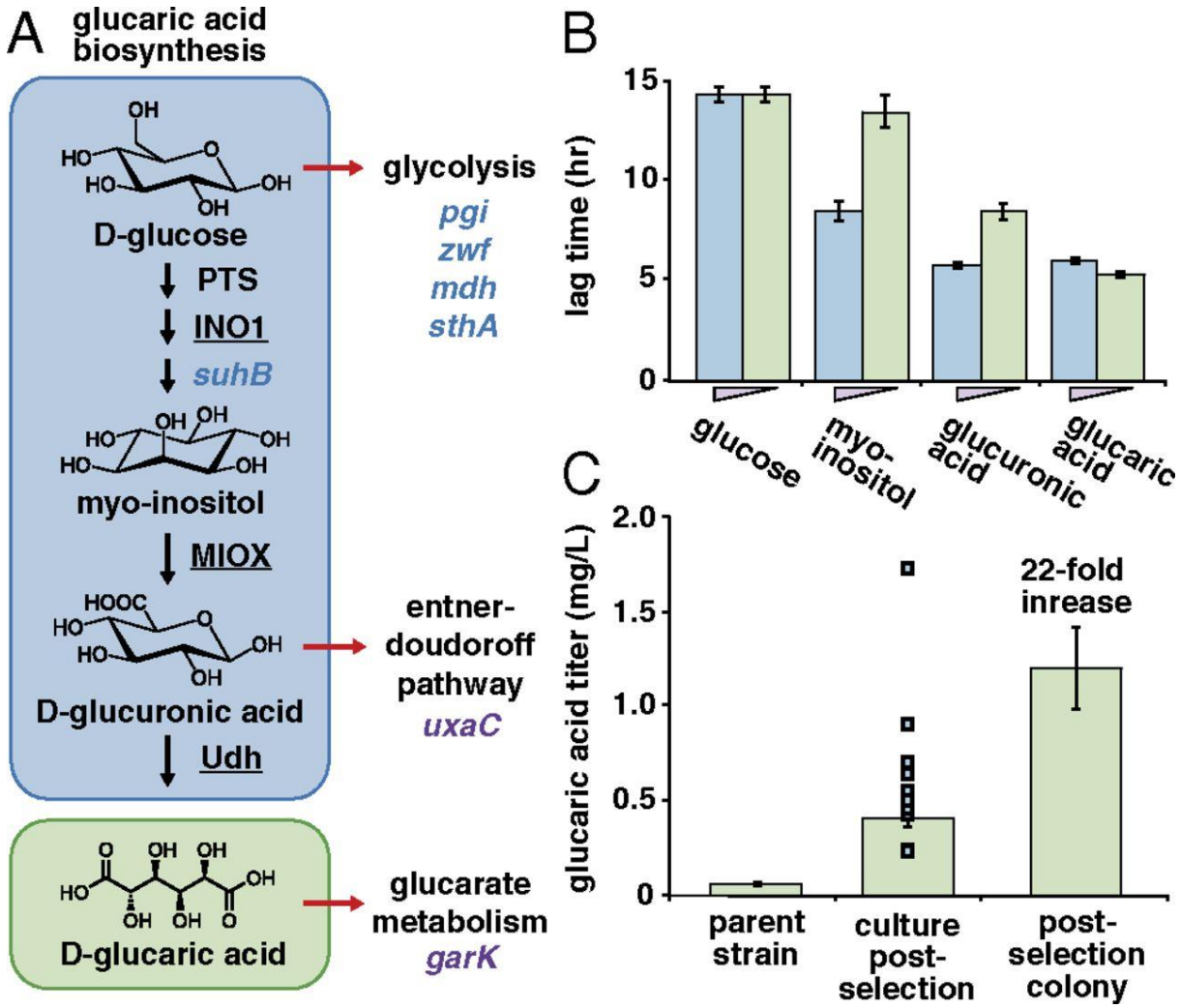


Figure 2.4 Optimization of the glucaric acid biosynthetic pathway. (A) Glucaric acid biosynthetic pathway showing key intermediate metabolites and enzymes. Heterologous gene names are underlined. Endogenous E. coli genes targeted by MAGE for expression modification: blue, RBS modification; purple, knockout. (B) Lag time in growth reflects time required for the pathway enzymes to produce activating levels of glucaric acid in the sensor selector strain CdaR−TolC. Pathway intermediates are supplied exogenously (blue, 10 mM; green, 1 mM). Error bars represent SEM from three biological replicates C) Glucaric acid titers produced by the parent strain, the postselection mixed population, and the highest-producing clone (bars). Squares indicate titers produced by clones isolated from the postselection population. Error bars represent SEM from three biological replicates.

To ensure that the heterologous enzymes were functional and provided a growth advantage under selective conditions, we measured growth lag times in the cdaR-tolC sensor-selector strain after exogenously providing pathway intermediates (glucose, myo-inositol and glucuronic acid). Furthermore, we verified cdaR is specifically activated by glucaric acid, and does not respond to pathway intermediates myo-inositol and glucuronic acid (Supplementary Figure S2.7).

Increasing concentrations of glucaric acid result in lower lag times for cells grown in the presence of SDS. Under selective conditions, decreasing growth lag times reflect the decreasing number of enzymatic reactions required to produce glucaric acid for cdaR-tolC activation (Figure 2.4B). Higher concentrations of myo-inositol and glucuronic acid resulted in shorter lag times under selective conditions, but increasing glucose or glucaric acid concentrations in the media did not result in a growth advantage. In the case of glucaric acid this is expected as both 1 mM and 10 mM are above the operational range. With glucose, one possible explanation is that an increase in glucose in the media results in additional flux through glycolysis and central metabolism rather than increased flux through the glucaric acid pathway, which likely operates slower than glycolysis. The lag time observed at the high glucuronic acid concentration is comparable to the lag time observed with glucaric acid, supporting the previous finding that the Udh enzyme acts on a fast time scale when compared to the selection [16, 117]. A long lag time even at a high concentration of myo-inositol indicates that the MIOX enzyme is less efficient as reported in previous work [118].

Efforts to increase glucaric acid production in *E. coli* have focused on co-localization of pathway enzymes [119] and improving MIOX solubility [118]. To date, modifying endogenous *E.*

coli pathways has not been explored. We hypothesized that glycolysis and the pentose phosphate pathway were competing with Ino1 for glucose-6-phosphate (g6p), the branch-point for glucaric acid production. We used MAGE to introduce degeneracy in the RBS of genes involved in catabolism of g6p (Supplementary Table S2.5). We similarly targeted the RBS sequences of *mdh* and *suhB*, the endogenous phosphatase responsible for dephosphorylating myo-inositol-1-phosphate [120] (Figure 2.4A). Degeneracy in the RBS sequences allowed the selection to sample both up- and down-regulation of the genes. We hypothesized that tuning the rate of glycolysis would allow the glucaric acid pathway to compete for glucose more effectively while still facilitating robust cell growth. The product of the *pgi* gene shuttles g6p into glycolysis and its disruption has been shown to increase the intracellular pool of g6p [121], the substrate of Ino1. The growth defect of a *pgi* mutant can be rescued by overexpression of *sthA* [120] and thus *pgi* and *sthA* were chosen for simultaneous expression modulation. The other major pathway for g6p catabolism is the pentose phosphate pathway and is initiated by the product of *zwf,* which was also targeted for expression modification. To prevent flux diversion of the intermediate molecule glucuronic acid into the Entner-Doudoroff pathway, we targeted uronate isomerase (*uxaC*) for a knockout. To avoid catabolism of glucaric acid, we also targeted glycerate kinase (*garK*) for a knockout.

We performed 5 cycles of MAGE on seven genomic targets (Supplementary Table S2.5) to achieve a predicted prevalence of approximately $1x10^{-6}$ for strains incorporating mutations at all seven loci. The statistically most common strain contained a single mutation and was predicted to account for 40% of the cell population. After MAGE followed by toggled selection, the enriched non-clonal culture produced 7-fold more glucaric acid than the parent. The best

clone isolated from this population produced 22-fold more than the parent (Figure 2.4C; Supplementary Table S2.7). This highest producing strain contained a targeted nonsense mutation in *garK*, a gene not previously shown to enhance glucaric acid production. None of the other targeted genes were mutated, but an untargeted nonsense mutation in the L-glyceraldehyde 3-phosphate reductase gene (*yghZ*) was found. As an aldo-keto reductase, *yghZ* has fairly broad substrate specificity [122] and could be diverting carbon flux away from glucaric acid by reducing glucuronate to gluconate.

Glucaric acid titers were improved 22-fold over the parent strain; however, absolute production of glucaric acid remained substantially lower (1.2 mg/L; Figure 2.4C) than previously reported titers [16]. Moon, et. al. [16] carried out glucaric acid production in an *E. coli* B-strain (BL21), while we optimized the pathway in the MAGE-competent E. coli K-strain. To investigate the possible role of strain background (B vs K strains) in glucaric acid production, we measured glucaric acid titer in our parent K-strain and BL21. We found that glucaric acid titer was 300 times higher in BL21 with the same glucaric acid enzymes and culture conditions (Supplementary Figure S2.5).

There are substantial differences between B and K strains of *E. coli* that are difficult to bridge through naïve mutagenesis. Notably, B strains have altered carbohydrate metabolism when compared to K strains as well as an enhanced capacity for recombinant protein production [117]. Previous work to produce glucaric acid in *E. coli* has revealed MIOX to be a highly unstable enzyme [16] and the primary limit on production may lie in protein folding and stability, rather than host-cell glucose metabolism. Our evolved K strain grew just slightly worse than the parent strain, ruling out gross metabolic deficiency as the cause of low production

(Supplementary Figure S2.8). In light of these considerations, subsequent rounds of diversification and selection were not pursued in the K-strain background. Currently, work is underway to enable MAGE in BL21 for optimization of production pathways better suited for *E. coli* B-strains. These results highlight that directed evolution is not a replacement for the careful choice of a host strain, but should complement thoughtful strain selection.

Discussion

Rapid advances in DNA sequencing and DNA synthesis technologies [40, 123] have not been accompanied by similar advances to enable the high throughput evaluation of phenotypes. Our implementation of small-molecule sensors coupled to selection advances a versatile platform that can transform biosynthetic phenotypes into fitness differences. These differences empower us to employ evolution followed by sequencing to reveal clues to potential metabolic pathway inefficiencies and to identify targets for subsequent rounds of evolution. The multiplex mutations facilitated by MAGE enable us to target all candidate genes predicted by FBA without prior assumption of the relative importance of each target. Because selection amplifies faster dividing cells, we indirectly enrich for variants that suitably balance biomass and biosynthetic objectives. We show that toggled selection refreshes the pool of productive cells by removing evolutionary escapees. Toggled selection enables multiple rounds of evolution to progressively enrich for higher producing variants. Combining beneficial mutations from independently evolved strains could lead to even higher metabolite production due to epistatic synergies. The incidence of evolutionary escapees and off-target mutations is likely to be significantly reduced by transiently repressing mismatch repair [124]. While this

may decrease untargeted beneficial mutations (e.g. mhpD, entB and hcaT in naringenin biosynthesis) in a single round of evolution, mutations that provide significant selective advantage will ultimately be enriched over multiple rounds.

Besides pathway optimization, we can use sensor-selectors to screen libraries of synthetic or metagenomic sequences for novel biosynthetic operons, new enzyme functions, and transporters. The vast reservoir of natural chemicals found in microbial species remain largely inaccessible because the enzymatic pathways for their synthesis are not known. With sensor-selectors, large libraries encoding natural or synthetic operons can be interrogated to identify the putative pathway for a target chemical.

Natural sensor domains exist for many classes of molecules that are of economic interest; however, some metabolite targets have no known sensor to detect them. We expect this challenge to be addressed by advances in protein design and by efforts to characterize new transcription factors encoded in metagenomes. Clever use of existing sensors will also allow the optimization of multiple pathways that use common intermediates. For biosynthetic pathways diverging only in late 'decoration' steps, we can leverage class-specific sensors to optimize the production of many related molecules by simply exchanging terminal enzymes. For example, our best naringenin production strain likely has an elevated intracellular concentration of malonyl-CoA, which could be used immediately for the improved production of fatty acid-derived targets or polyketides.

Evolution is a powerful tool for resolving the complexity of biology. Using evolution to guide rational design should ultimately lead to a better understanding of the genotypic basis of biological function.

Methods

Riboswitch sensors

Because riboswitches affect expression control through translation of the transcript rather than transcription, a modified sensor architecture was used. The theophylline-responsive theoRR blocks translation unless theophylline is present [70]. We included the theoRR as a 5' untranslated region of the *tolC* transcript, allowing for translation only when theophylline was present.

The *btuB* vitamin B12-responsive riboswitch operates in an opposite manner to theoRR, remaining in an open conformation natively, and attenuating translation through RBS occlusion only when B12 or its derivatives are present [48]. We included the *btuB* riboswitch and the first 70 codons of *btuB* gene sequence [48] at the beginning of the tetR transcript in a modified tetR-CAT sensor-selector. When the btuB-tetR fusion is transcribed in the absence of vitamin B12, the riboswitch structure within the btuB 5'-UTR is in its open conformation, allowing translation of the btuB-tetR fusion protein, which represses the transcription of CAT, leaving the cells chloramphenicol sensitive. In the presence of vitamin B12, the riboswitch changes conformation to bind B12 and occludes the RBS, preventing btuB-tetR translation and allowing strong transcription of CAT, leading to chloramphenicol resistance. Using this strategy, the translation-attenuating B12-responsive btuB riboswitch was used to control the expression of a positive selection marker.

Sensor-selector strain construction

To construct the TtgR-TolC dual sensor strain, a linear PCR product composed of 5'-zeocin resistance cassette—apFAB101 promoter [7]—RBS B0034 (aaagaggagaaatta)—*ttgR*-3' was amplified with 50 bp homology and recombineered into the genome of *E. coli* K12 derivative EcNR2 [30] at locus 1529620 (numbered relative to the MG1655 sequence) of the TtgR-TolC sensor-selector strain. (For pathway optimization, the pLtetO promoter in the *ttgR* sensor gene construct at the *tolC* locus was also replaced with apFAB101 promoter to avoid homology with pLtetO promoters on naringenin pathway plasmid 2. To construct TtgR strain with dual selectors, a linear PCR product composed of 5'- spectinomycin resistance cassette—*ttgA*p promoter -3' was amplified with 50 bp homology and recombineered 5' to the CAT gene at the Δ*mutS::Cm* locus of the into the TtgR-TolC sensor-selector strain.

Riboswitch-based sensor-selectors necessitated modifications to the standard construction used for all transcription factor-based sensor-selectors. The theophylline-responsive riboswitch (theoRR) sensor used promoter pLtetO to directly transcribe *tolC*, with the theoRR appended 5' to the *tolC* start codon as a 5' untranslated region. The *btuB* riboswitch invertor was a modified version of the TetR-CAT sensor-selector. The *tetR* gene ATG start codon was replaced by the 239 bp of the *E. coli btuB* 5'-untranslated region, and the first 210 nucleotides of the *btuB* coding sequence [48].

Sensor-selector strains CdaR-TolC and dual TtgR-TolC were modified before pathway optimization to include T7 RNA polymerase under control of a pLac promoter. A linear PCR product containing 5'- spectinomycin resistance cassette reverse complement—pLac—T7 RNA

polymerase gene - 3’ was amplified with 50 bp homology and recombineered into the genome to replace the *bla* gene within the Red-λ prophage.

Escape rate measurements

Sensor-selector escape rates were measured by growing cells overnight to saturation in LB medium. Serial dilutions of cells were plated onto non-selective plates to measure the saturation culture density (colony forming units/mL). Serial dilutions of cells were plated onto appropriate selective plates (LB supplemented with SDS or chloramphenicol) to measure the density of cells (CFU/mL) surviving selection in the absence of chemical inducers. The density of cells surviving selection was divided by the total saturation density to calculate the escape rate of each sensor-selector.

TtgR-TolC sensor-selector degradation tags modification

We appended three *ssrA* degradation signal variants of varying strengths [109, 125] to the 3’ end of the *tolC* selector gene coding sequence in the ttgR-tolC sensor-selector strain. The ssrA variants were inserted in frame to replace the stop codon of the selector by recombineering, using a zeocin resistance cassette as a selection marker for integration. The following degradation tags were appended to the selector gene in frame: strong (RPAANDENYALAA*), medium (RPAANDENYAAAV*), and weak (RPAANDENYALVA*). Each strain was sequence verified. We determined the escape rate and operational range for each degradation tag using orthogonal gradient growth assays.

TtgR-TolC sensor-selector RBS modification

We generated ten RBS variants by increasing or decreasing the separation between the Shine Dalgarno sequence of promoter *ttgA*p (5'- CCGAGGATCCTC -3') and *tolC* translation start site by 1 to 5 bases (Supplementary Figure S2.6). The underlined bases remained unmutated. We designed oligonucleotides for each RBS variant and used MAGE [30] to modify the TtgR-TolC sensor-selector strain. We verified the ten variants by sequencing colonies after MAGE, and determined the escape rate and operational range for each of the ten RBS variants using orthogonal gradient growth assays. We measured operational range and escape rates of the five RBS variants that showed ligand induction response (Figure 2.2B; Supplementary Figure S2.6).

TetA exporter plasmid construction and assay conditions

Plasmid pKD46 [126] (Genbank accession AY048746) was modified, replacing lambda red *exo*, *beta* and *gam* genes with the *Streptococcus* tetracycline efflux pump (*tetA*) gene, and also replacing beta lactamase (*bla*) gene with spectinomycin resistance gene to create plasmid pBAD-tetA. The *tetA* gene is transcriptionally controlled by the arabinose-inducible pBAD promoter. Plasmid pBAD-*tetA* was transformed into TetR-CAT sensor-selector strain. This strain was characterized using orthogonal gradient growth assays under the following *tetA* expression induction conditions: no arabinose, 0.05 % arabinose (intermediate), or 0.5 % arabinose with 4 hr pre-induction (high).

Orthogonal gradient growth assay

Arrays of growth conditions were evaluated in 96-well plates (BD-Falcon). Serial dilution of selection agent (SDS, chloramphenicol or colicin E1) along one axis of the 96-well plate was followed by serial dilution of an inducer chemical along the second axis to create the arrays in 150 µl of kanamycin supplemented LB media. Overnight cultures picked from fresh colonies were inoculated at a dilution factor of at least 1000 into each well. Plates were incubated at 30 °C with agitation and measurement of optical density at 600 nm (OD600) for at least 16 hours in a Biotek (Winooski, VT) plate reader.

Chemical detection concentration and saturation concentration thresholds were found from these assays by determining the minimum chemical concentration at which cells showed a growth response when compared to no chemical (detection threshold) and the concentration at which higher chemical concentration showed no additional growth benefit (saturation threshold). Data from single measurements are reported in Figure 2.2A.

Glucaric acid production conditions

Overnight cultures of cells were diluted 1:100 into selective LB supplemented with 50mM glucose and 1mM IPTG. 2ml deep 96-well blocks with a working volume of 1ml were incubated humidified at 37C and 900RPM. After 72 hours the plates were spun down for 5 minutes at 4000rpm and the supernatants were filtered prior to analysis.

Naringenin production conditions

Cell were grown in M9 minimal medium supplemented with 1 % glucose, 1 mM biotin, kanamycin and carbenicillin for 24 hours at 30 °C in deep 96-well plates with fast shaking (900

rpm). After 24 hours, cultures are diluted 1:20 into fresh M9 glucose medium with 1 mM IPTG and incubated for 72 hours in deep 96-well plates with fast shaking.

Glucaric acid LC-MS analysis

Detection and quantitation of glucaric acid from culture supernatant was achieved by LC/MS/MS analysis performed with an Agilent (Agilent Technologies, Santa Clara, CA) 6460 triple-quadrupole LC/MS/MS system. The ([M - H+]- / product ion) monitored via electrospray ionization in the negative ion mode with multiple reaction monitoring (MRM) was (209/85.1 amu). Mass spectrometer parameter settings were gas temp (350 °C), gas flow (12 L/min), nebulizer pressure (35 psi), sheath gas heater (400 °C), sheath gas flow (12 L/min), and capillary (4000V). An external standard curve mixture was analyzed at various concentrations 20 pg/uL – 10 ng/uL and utilized for quantitation.

Hydrophilic interaction chromatography (HILIC) conditions with a Phenomenex Luna 5u NH2, 250 x 2.00 mm column (Torrance, CA) were as follows: flow rate: 0.4 mLs/min; solvent A: 20 mM ammonium acetate + 20 mM ammonium hydroxide in 95:5 water:acetonitrile, pH 9.45; solvent B: acetonitrile. The gradient was as follows: *t* = 0, 85 % B; *t* = 10 min, 0 % B; *t* = 11, 0 % B; 3 minute equilibration.

Naringenin LC-MS analysis

Naringenin was extracted from culture supernatant by mixing with an equal volume of ethyl acetate, mixed vigorously for 30 seconds on a benchtop vortexer, and the mixture briefly centrifuged for phase separation. The organic (upper) phase was transferred to a glass vial, and

the ethyl acetate evaporated by gently circulating inert gas into the glass vial. The sample is resuspended in 200 uL of methanol and stored at -20 °C.

Samples were analyzed by LC/MS using a Bruker maXis impact Q-Tof (Billerica, MA) with an Agilent 1290 HPLC (Palo Alto, CA). A 2 x 100 mm Phenomenex Gemini column with 3 µm particle size (Torrance, CA) was used for a gradient separation. Mobile phase A was 0.1 % formic acid in water and mobile phase B was acetonitrile with a flow rate of 0.2 mL/min. The gradient began with 0 % B for 1 min, then increased to 100 % B after 12 min and held for 3 min before returning to 0 % B. Total run time was 20 min. The mass spectrometer was operated in the MRM mode selecting $(M-H)^-$ for coumaric acid (*m/z* 163) and naringenin (*m/z* 271) with a collision energy of 20 eV for each. The total ion chromatograms from each product ion scan were plotted separately and integrated. Naringenin standards at 2 and 10 µM were used as standards.

Whole genome sequencing

Strains for whole genome sequencing were grown overnight to saturation in LB supplemented with appropriate antibiotics. Genomic DNA was extracted from 1 ml culture using the Qiamp DNA Mini kit (Qiagen, Hilden, Germany). 1 µg genomic DNA was sheared using a Covaris S2 Ultrasonicator (Covaris, Woburn, MA). Paired-end libraries were prepared using the TruSeq DNA kit (Illumina, San Diego, CA). After quantitative PCR quantification, libraries were sequenced using paired end 300 bp sequencing on a MiSeq (Illumina) to mean of >100 coverage depth per genome. Reads were aligned to the MG1655 genome sequence, and single nucleotide polymorphisms (SNPs) and structural variations were detected using Lasergene SeqMan Pro 11 software (DNAStar, Madison, WI).

Bioreactor production of naringenin

We used an Eppendorf Celligen 310 bioreactor for fed-batch fermentation of the highest producing naringenin strain from the fourth round of evolution. A 1 liter culture of the strain was grown in M9 medium supplemented with 1 % glucose, casamino acids and biotin for 5 days at 32 °C with constant flow of pressurized, filtered air. The pH of the culture was buffered with ammonium hydroxide, and foaming was suppressed with anti-foaming agent. Glucose and casamino acids were provided every 12 hours. Samples were drawn periodically and assayed for coumaric acid and naringenin concentrations.

Chapter 3. Engineering an allosteric transcription factor to respond to new ligands

Attributions

The majority of this chapter was previously published in Nature Methods [93] and does not require explicit permission to reproduce here.

Noah D. Taylor, Farren J. Isaacs, George M. Church and Srivatsan Raman conceived the study. Noah D. Taylor, Stanley Fields, George M. Church and Srivatsan Raman designed experiments. Noah D. Taylor, Alexander S. Garruss and Srivatsan Raman performed experiments and carried out bioinformatic studies. Rocco Moretti and David Baker generated computational protein design candidates. Sum Chan, Duilio Cascio, Mark A. Arbing and Sriram Kosuri solved the crystal structure of a sucralose-binding variant. Sriram Kosuri helped with Agilent OLS chip library design. Jameson K. Rogers helped optimize screening protocols. Noah D. Taylor, Alexander S. Garruss, Stanley Fields, George M. Church and Srivatsan Raman analyzed the data. Noah D. Taylor, Alexander S. Garruss, Stanley Fields, George M. Church and Srivatsan Raman wrote the paper.

Summary

Genetic regulatory proteins inducible by small molecules are useful synthetic biology tools as sensors and switches. A major class of regulatory proteins is microbial allosteric transcription factors (aTFs), but aTF-inducer pairs are currently limited by those that naturally occur. Altering inducer specificity in these proteins is difficult because mutations that affect

inducer binding may also disrupt allostery. Here, we engineer an aTF, LacI, to respond to one of four new inducer molecules: fucose, gentiobiose, lactitol or sucralose. We employ computational protein design, single-residue saturation mutagenesis, or random mutagenesis, along with multiplex assembly, and identify initial hits via a two-stage enrichment screen. Following activity maturation, we identify LacI variants with specificity to and induction by these new inducers comparable to that of wild-type LacI and its inducer, IPTG. The ability to create designer aTFs will enable applications including dynamic control of cell metabolism, cell biology and synthetic gene circuits.

Introduction

Allosteric transcription factors (aTFs) encompass several large families of proteins that provide environmental response in bacteria. Upon binding a small molecule, aTFs undergo a conformational change that alters their affinity for an operator DNA sequence that is often found upstream of metabolic operons or transporter genes [52, 55, 106, 107]. Allosteric transcription factors have been co-opted for use as gene expression switches [45] which form a cornerstone in synthetic biological applications. For example, aTFs can serve as intracellular metabolite sensors to enable directed evolution of biosynthetic pathways [43, 47], as devices to control information flow and feedback regulation in synthetic gene networks [127], and as switches in metazoan systems to provide synthetic control of cell differentiation and development [79, 128].

Expanding aTFs to respond to new molecules would greatly increase their utility, facilitating the biosynthetic production of valuable natural products, fine chemicals and

intermediates, and fuels [2, 129]. However, inducer recognition and transcriptional response in aTFs are tightly coupled through allostery, making redesign toward new inducers challenging. Residues mediating allostery are generally unknown, but can be distributed throughout the protein structure [130], and mutations to the ligand-binding site often disrupt allosteric communication with the DNA-binding domain [131, 132]. New high throughput genetic approaches offer the possibility of understanding allostery at molecular resolution (see Chapter 4) [133], but this promise has yet to be realized. Previous efforts to engineer aTFs have shown that random mutagenesis and screening can modify aTF proteins to respond to inducers closely related to their wild-type (WT) inducer [61, 63, 64, 134-136]. These include introducing inducer specificity to LuxR, a promiscuous aTF [63, 64], or broadening the inducer profiles of XylR [136] and AraC [134, 135]. However, the limitations of random mutagenesis preclude applying this approach toward an arbitrary ligand.

Here, we present a general strategy to engineer an aTF to respond to new inducer molecules, using the *E. coli lac* repressor, LacI, as a test case. We evaluate thousands of candidate variants, derived from computational design, protein-wide single amino acid substitution or error-prone PCR, by a two-stage screen that sequentially enriches for variants that are both allosterically functional and responsive to a new ligand. We mature the activity of these initial hits toward greater specificity and stronger induction. We demonstrate the utility of this approach by engineering LacI variants to respond to gentiobiose, fucose, lactitol or sucralose with response comparable or superior to the wild-type LacI response to its synthetic inducer, isopropyl β-D-1 thiogalactopyranoside (IPTG). Though the LacI protein is well-

characterized, we incorporate no *a priori* information on its allosterically important residues, using knowledge only of its structure and DNA operator binding site.

Results

Choice of new inducer molecules

LacI, which natively regulates the lactose catabolism operon, *lacZYA*, in response to the disaccharide allolactose, also responds to IPTG. As new inducer targets, we chose four saccharides and glycosides not metabolizable by *E. coli*: gentiobiose, fucose, lactitol and sucralose [137]. These molecules represent targets with increasing apparent chemical difference from known LacI inducers (Supplementary Figure S3.1A). Lactitol and sucralose are synthetically-derived chemicals.

Design, synthesis and assembly of candidate variants

To capture the effects of protein residues both proximal and distal to the ligand-binding pocket, we used three methods to create LacI variants: computational design, protein-wide single amino acid saturation mutagenesis, and error-prone PCR.

LacI variants were computationally designed using an adaptation of the Rosetta software suite [26, 27] to bind the three target ligands – fucose, lactitol and sucralose – most dissimilar from the native inducer (see Supplementary Methods for Chapter 3). Rosetta has been successfully used to design proteins with new ligand binding interactions [92], though it cannot account for allostery. DNA oligonucleotides encoding designed variants were generated by microarray-based synthesis [39, 40], which specifies a pool of exact oligonucleotide sequences. Due to oligonucleotide length limitations, residues mutable during Rosetta design

were confined to three segments of *lacI* (encoding residues 73-125, 148-197 and 245-296), encompassing the majority of the binding pocket. LacI libraries encoding each single segment (mean 4.2 mutations per gene) were synthesized and cloned, and were combined through overlap PCR to capture full designs with mutations in each segment (mean 12.6 mutations per gene).

Mutations to LacI residues distal to the ligand interface can influence induction through long-range effects [138-140]. Thus, we created a variant library encoding all LacI single amino acid substitutions using microarray-synthesized DNA (CustomArray) by tiling mutable sequences in windows of 36 residues, totaling 6,800 variants. Sampling by next generation sequencing (MiSeq, Illumina) indicated that this library captured ~88% of all single mutations, with ~74% of the positions encoding at least 17 of the 19 possible substitutions (Supplementary Figures S3.2B and S3.2C). Finally, LacI codons 67-197 were amplified by error-prone PCR mutagenesis (GeneMorf II, Agilent Technologies) to generate a library with a mean of 5 mutations per gene.

A screen to identify LacI variants with new ligand responses

Screening methods such as phage or yeast display or two-hybrid screening can evaluate binding, but not allostery, so we developed an *in vivo* selection-screening method designed to capture aTF variants functional in both allosteric states: DNA-bound in the absence of inducer, and allosterically activated by inducer (Figure 3.1A). Into the genome of *E. coli*, we integrated a single copy of a reporter construct consisting of the genes encoding green fluorescent protein (GFP) and TolC under transcriptional control of the LacI-regulated promoter, pLlacO [45]; TolC is an *E. coli* outer membrane porin that mediates the entry of the bacteriocin toxin, colicin E1 [31,

111]. LacI variants that bind DNA and repress transcription were first enriched by colicin E1 selection (Figure 3.1B), and subsequently, variants that activate transcription in response to a target ligand were collected by fluorescence-activated cell sorting (FACS; Figure 3.1C). GFP-positive cells were then assayed clonally by high-throughput flow cytometry to measure baseline GFP and induction ratios (fluorescence ratio of induced to uninduced cells) for the new inducer (Supplementary Table S3.1). Wild-type LacI is induced 15-fold by IPTG in this screening system (Supplementary Figure S3.4).

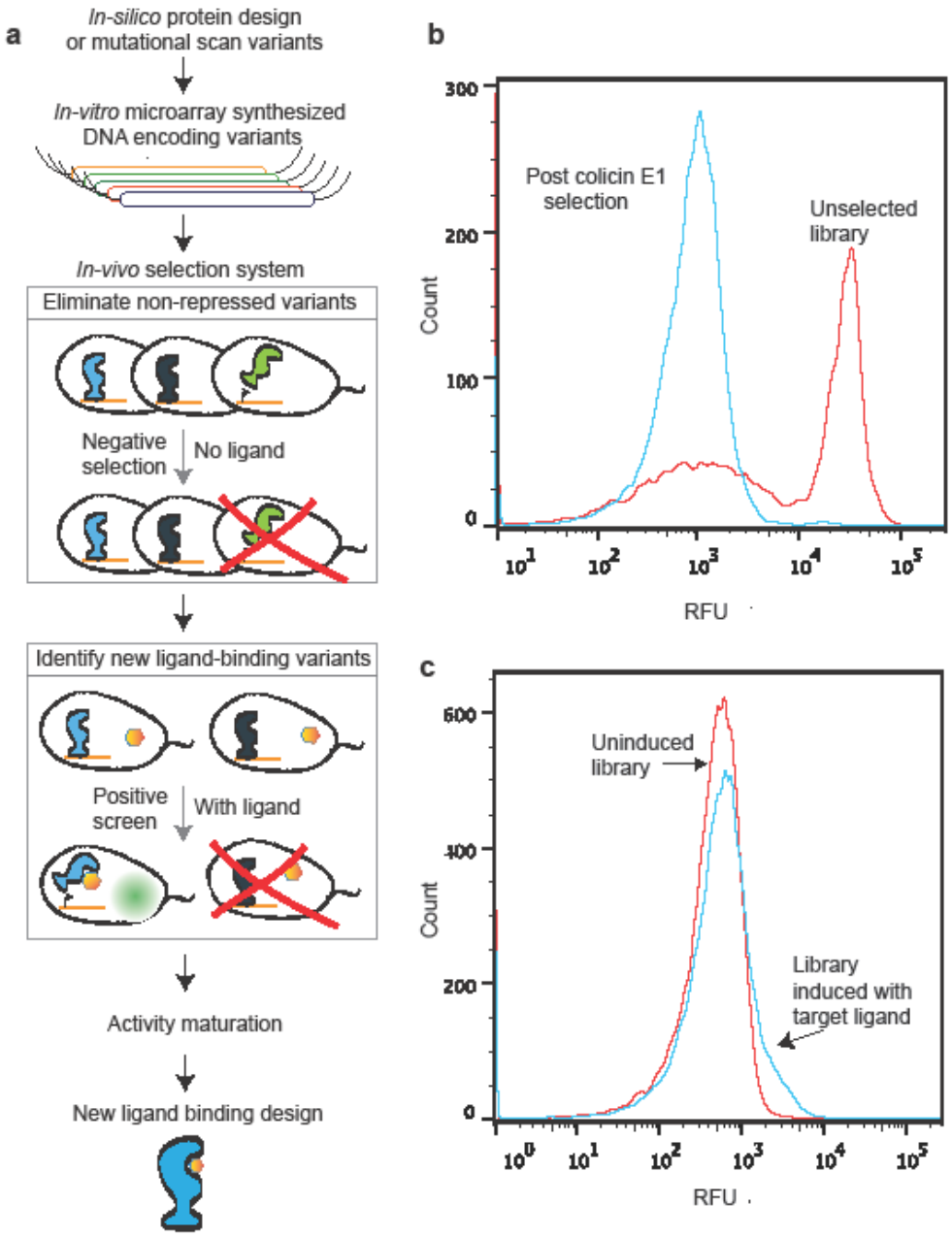


Figure 3.1. General workflow for designing new ligand binding in an allosteric transcription factor. (a) Schematic diagram showing the steps in the design workflow: computational protein design; synthesis of oligonucleotides encoding the designed variants; a two-step screening system for eliminating variants incapable of repressing transcription, and subsequently selecting variants that respond to a new ligand; followed by activity maturation. (b) Fluorescence histogram of E. coli cells containing LacI variants before (red) and after (blue) colicin E1 selection (no ligand present). Cells containing variants that do not repress DNA show higher fluorescence (right side) compared to those that repress DNA (left side). (c) After colicin E1 selection, fluorescence histogram LacI variants after incubation with target ligand (here gentiobiose). The leading edge on the right side contains the ligand-responsive variants.

Instead of the more commonly used multicopy plasmid-based reporter, we genomically integrated a single reporter copy to minimize fluorescence artifacts arising from fluctuations in plasmid copy number. Higher reporter copy numbers allow higher total fluorescence which yields higher fold induction [44]; in a multicopy reporter system, the induction of each LacI variant would be expected to scale accordingly, higher than the values reported here. Although GFP has been used as a positive and negative screen reporter, as shown by Tang *et. al*. [135], we preferred a TolC-based negative selection for enriching transcriptionally-repressed variants owing to the poorer resolvability of the flow cytometer at low fluorescence levels. We verified that colicin E1 selection strongly enriched (> 99.5%) for clones encoding full-length *lacI* sequences devoid of frameshifted variants, a common occurrence due to deletion errors in array-synthesized DNA.

Computational design was targeted to three segments of the LacI protein (residues 73-125, 148-197 and 245-296), which form the ligand binding pocket. Libraries of designed mutants within each segment were synthesized and cloned into the LacI expression plasmid. When the three libraries were combined to build full Rosetta designs through overlap extension PCR, a much greater fraction (> 98%) of variants than expected from synthesis errors alone (38%) failed to repress GFP expression. Since Rosetta designed only for ligand binding, we hypothesized that many Rosetta designs suffer from a high mutational burden that inactivates allostery; indeed, 14 of the 15 highest ranking Rosetta designs (five per ligand) failed to repress transcription when tested independently (Supplementary Figure S3.1B).

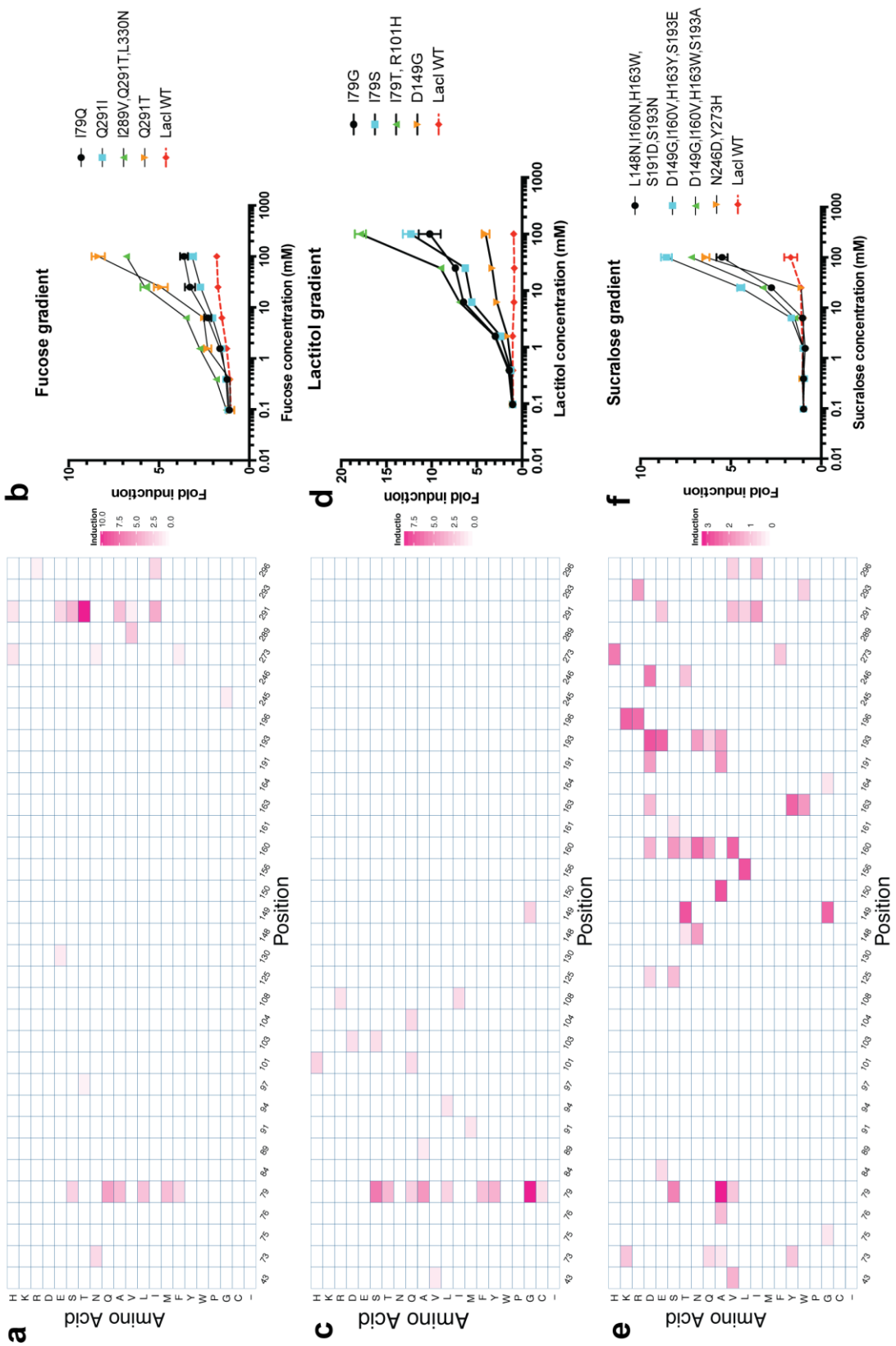
We therefore chose to screen the libraries of single Rosetta-designed segments for target ligand response. In addition, we screened LacI single amino acid substitution and error-prone PCR libraries for induction by each ligand.

New ligand responses by LacI variants

We identified Rosetta-designed LacI variants that responded to fucose, lactitol, or sucralose (Figure 3.2). The best clones showed induction values similar to WT LacI induction by IPTG (15-fold; Supplementary Figure S3.4), demonstrating that new ligand binding can be engineered without compromising allosteric regulation. For each target ligand, we found multiple distinct single mutations or combinations of mutations that resulted in response to the same ligand (Figure 3.2; Supplementary Figure S3.2; Supplementary Table S3.1). The diversity of responsive mutants varied across the three ligands. For example, sucralose-responsive sequences were the most diverse, with the most responsive clones often containing four or more mutations (Figure 3.2G-3.2I; Supplementary Table S3.1). Fucose response was mediated by independent mutations to residues in different regions of the binding pocket (Q291T or I79Q; Figure 3.2A-3.2C), but lactitol response nearly always required a mutation to I79 (Figure 2D-2F). Mutations to I79 and Q291 were frequently present in variants that responded to multiple inducers (Figs 3.2B, 3.2E, and 3.2H), suggesting that these residues may determine ligand specificity of the binding pocket.

Figure 3.2. Characterization of Rosetta design variants responding to new inducers. (a-c) fucose (d-f) lactitol and (g-i) sucralose. The left side of each panel shows the ligand-binding domain of a LacI monomer (PDB ID: 2P9H) with positions found mutated in ligand-responsive variants colored in red. The center panel shows amino acid substitution profile at each position with the color shade denoting fold induction response. The highest fold induction, adjusted for number of mutations, is used for each amino acid substation. For example, with the lactitol-responsive variants, I79T shows the greatest induction (4.36) when present with R101H but gives an induction of 3.44 on its own. The right panel shows a dose response curve of four different variants responding to their target ligands (solid lines). WT LacI dose response is also shown (red dashed lines).

Figure 3.2 (Continued)



The I79 substitutions found in the top lactitol-responsive variants tended to be small (Gly or Ala) to accommodate the bulky ligand, or have hydroxyl- or thiol-containing sidegroups (Thr, Tyr or Cys) capable of hydrogen bonding with the sugar alcohol, and were mostly distinct from the substitutions in fucose-responsive I79 variants (Gln, Ala, Met or Leu). This comparison suggests that I79 plays an important role in ligand specificity. However, computational studies predicting ligand specificity-determining residues in LacI family proteins failed to identify either I79 or Q291 as key determinants of specificity [141-143]. Thus, our structure-based computational design targeted cryptic determinants of ligand specificity. Furthermore, we compared our laboratory-evolved fucose-responsive variants to naturally-occurring fucose-responsive aTFs in the GalR/S family. We found that three designed mutations conferring fucose response (I79L, I79M, and 273F; Supplementary Figure S3.5) were also differentially conserved between GalR/S and orthologous LacI sequences (P = 0.0471). This result suggests that our design method recapitulates natural evolutionary solutions to fucose binding.

We tested the utility of error-prone PCR for aTF design because this method is widely accessible with common laboratory reagents and does not require computational resources for protein design or a DNA synthesizer. Error-prone PCR generated variants responsive to fucose or lactitol but not sucralose, and was much less effective than computational design as measured by the maximum induction of variants (10.5- vs. 5.0-fold for fucose, and 7.1- vs. 4.8-fold for lactitol, Supplementary Table S3.2) or the proportion of variants after sorting that showed 2-fold or greater induction (42.7% vs. 17.7% for fucose, and 27.1% vs. 5.2% for lactitol, Supplementary Table S3.2). The computationally-designed variants with the strongest response to sucralose contained four mutations. This combinatorial complexity of mutations is likely too

large to be sufficiently sampled by error-prone PCR, explaining the absence of sucralose-responsive variants within the error-prone PCR library.

Error-prone PCR failed to generate any sucralose-responsive clones, despite the discovery that the computationally-predicted single mutants I79A and L196K show moderate response to sucralose (Supplementary Table S3.1). Despite there existing just 6859 possible LacI single residue mutations, failure to generate either of these sucralose-responsive mutations becomes less surprising when the Hamming distances are examined. Hamming distance in this context measures the minimum number of substitutions required to change from one codon to another (i.e. mutational distance). Codon 79 of *lacI* encoding isoleucine (ATT) lies a minimum DNA Hamming distance of two from all possible codons encoding alanine (CNA), as does codon 196 encoding leucine (CTG) from both possible codons encoding lysine (AAR; Supplementary Figure S3.7). For *lacI*, a gene of length 1083 basepairs, there are 3.5×10^6 possible DNA double mutations, making either of these specific mutants quite rare in a randomly mutagenized library. By utilizing microarray-synthesized oligonucleotides, this significant limitation of random mutagenesis can be avoided, as every single residue mutation lies a protein Hamming distance of one apart. Arguably, any arbitrary set of mutations within a region encodable on a single oligo lies an effective Hamming distance of one from every other set, illustrating the power of microarray-printed DNA.

Distributed mutations affecting LacI ligand binding

Mutations distal to the ligand- or DNA-binding domains influence the affinity of LacI for ligand or DNA through cryptic allosteric networks [138-140]. We systematically investigated this

effect on new ligand binding by assaying all single amino acid substitutions of LacI against gentiobiose, a close relative of the native inducer, allolactose.

From single amino acid saturation libraries, we identified many gentiobiose-responsive variants, including some with mutations far from the ligand-binding site; mutations primarily clustered into three regions: the binding pocket, the dimerization interface, and the DNA-binding domain (Figure 3.3A and 3.3B). The top variants with a single mutation in the binding pocket (Q291H), dimerization interface (R255H), or DNA-binding domain (V20A) showed similar induction (7.7-, 8.4-, and 6.7-fold, respectively), suggesting that allosteric effects of distal mutations are as potent as ligand-proximal mutations in the binding pocket (Figure 3.3A and 3.3B).

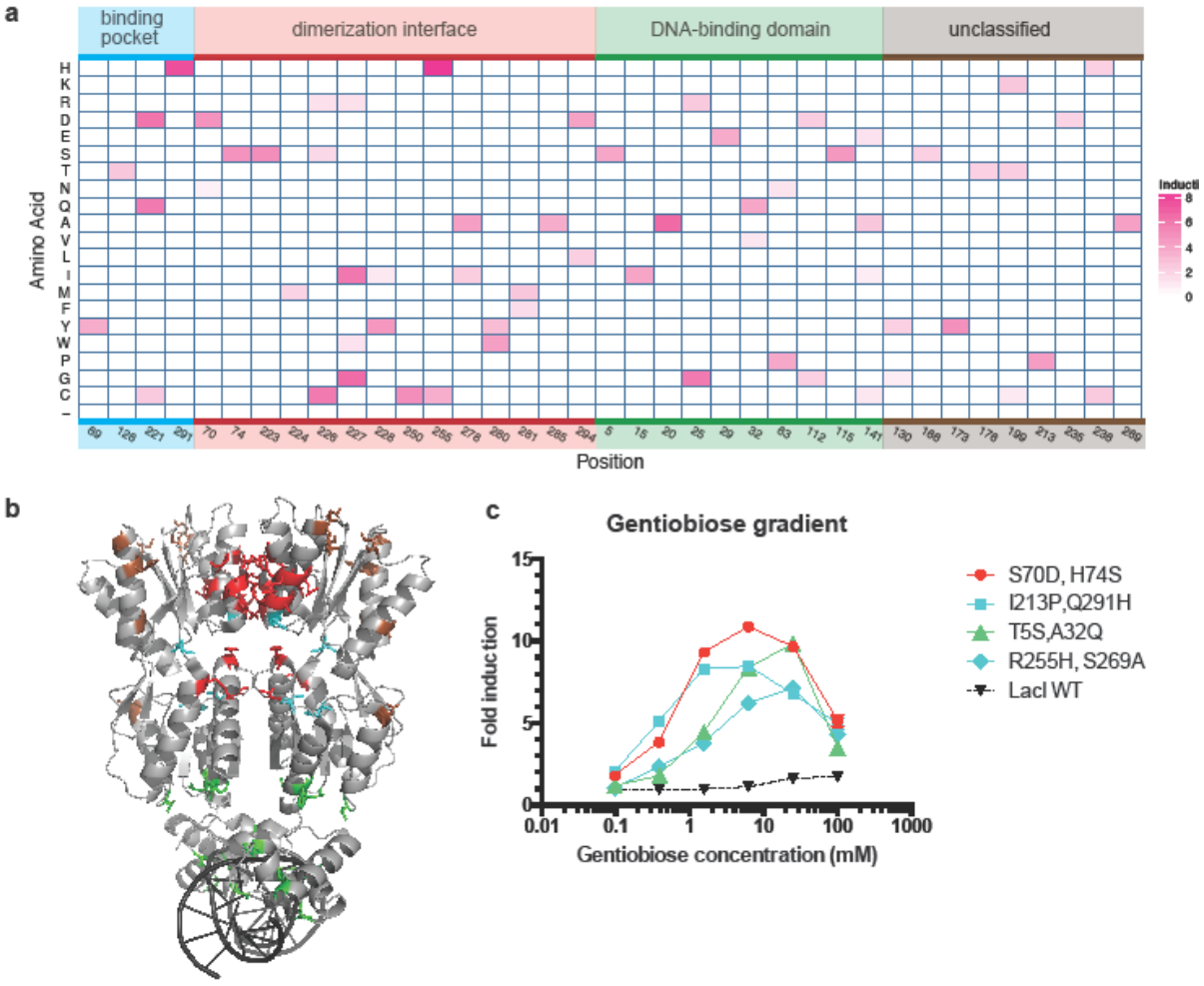


Figure 3.3. Characterization of gentiobiose-responsive variants from the protein-wide single amino acid substitution library. (a) Amino acid substitution profile at each position with the color shade denoting fold induction. Mutations in each variant is classified into four groups based on their location on the structure: cyan – ligand binding pocket; red – dimerization interface; green – DNA-binding domain; brown – not classifiable into any of the other three groups. (b) Residues found mutated in gentiobiose-responsive variants (>= 4.0 fold induction) are highlighted on the structure (PDB ID: 1LBG) with WT sidechains shown. The color scheme of different groups is the same as panel A. (c) Dose-response curves of four gentiobiose-responsive variants. Colors of the curves correspond to the structural classification of location of mutation. The dotted black line is WT LacI response to gentiobiose.

A large fraction of gentiobiose-responsive variants had mutations at the dimerization interface of the ligand-binding domain (Figure 3.3A). Despite our design of only single mutations, the most responsive variants each contained an additional mutation arising from DNA synthesis errors. The double mutant, S70D/H74S, with both mutations located in the dimerization interface of the N-terminal domain, showed the highest induction (> 10-fold; Figure 3.3C). Library sequencing before and after colicin E1 selection for DNA binding found that many mutations in the dimerization interface ablate DNA binding (Supplementary Figure S3.3A). For example, at residue A250, 12 of 19 possible mutations were not tolerated, but the permissible A250C mutation generated gentiobiose response (Supplementary Figure S3.3A). These results are consistent with mutational and biophysical studies that show that the allosteric signal in WT LacI upon IPTG binding involves communication between monomers via the dimer interface [144].

More surprisingly, mutations within the DNA-binding domain (e.g. T5S, V15I, V20A, N25G and H29E) or near this domain (e.g. H112D, H112G and L115S), about 40-50 Å from the ligand-binding pocket, retained DNA binding yet yielded a strong gentiobiose response (Figure 3.3A and 3.3B). We measured the dose-dependent response of four top variants with mutations within distinct regions: the binding pocket (Q291H/H173Y); N-terminal dimerization interface (S70D/H74S); C-terminal dimerization interface (R255H/S269A); and the DNA-binding domain (T5S/A32Q; Figure 3.3C). This positional diversity underscores the whole-protein phenomenon of allostery, suggesting that many distal mutations can subtly rearrange the conformation of the ligand-binding domain to alter ligand specificity.

Ligand promiscuity of LacI variants

We found that nearly all variants responsive to a new inducer retained a strong response to IPTG. To assess ligand promiscuity, we measured the induction of select variants against all five ligands: fucose, lactitol, sucralose, gentiobiose and IPTG (Figure 3.4). Ligand promiscuity was widespread, as most variants showed some reactivity to more than one new inducer.

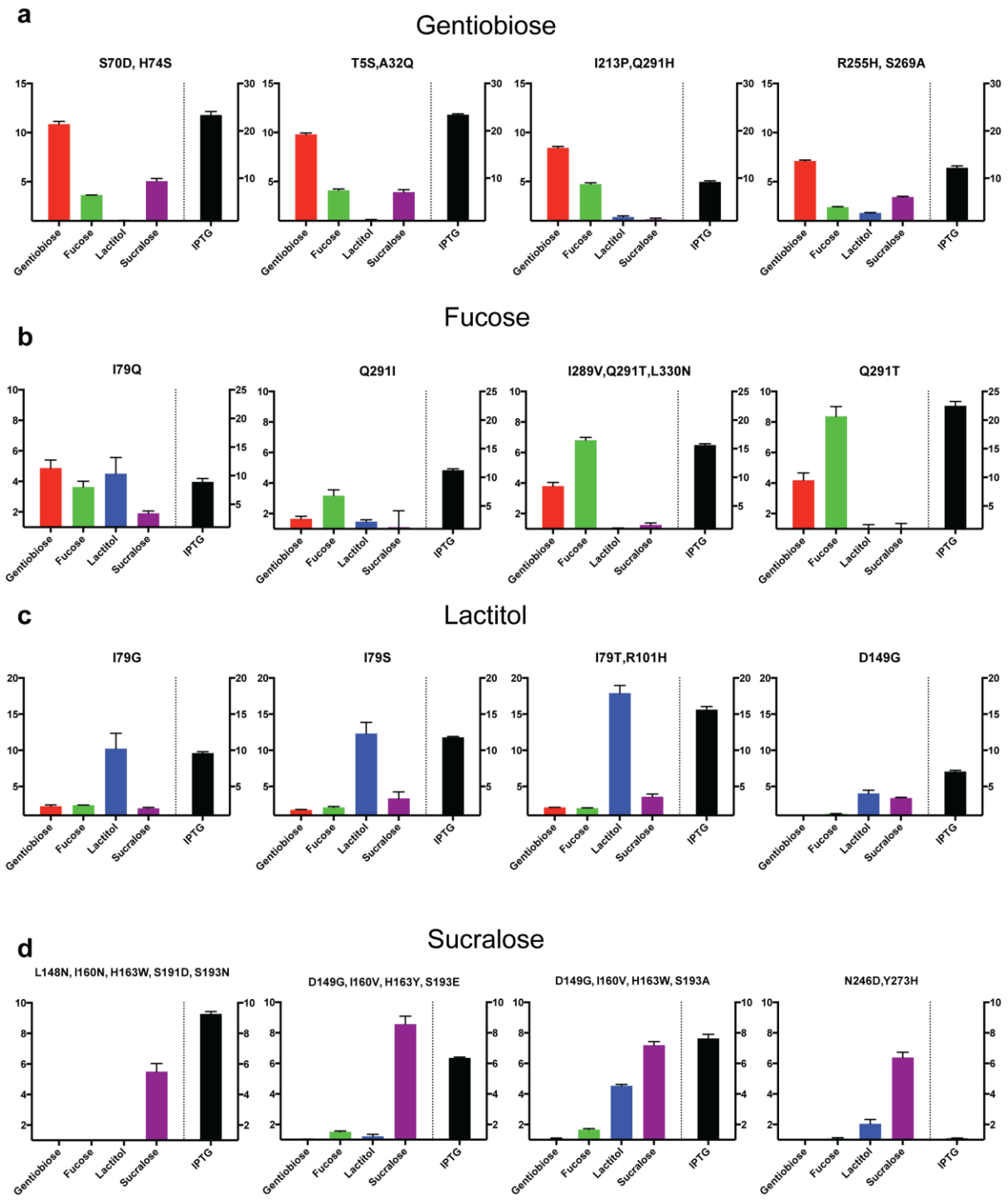


Figure 3.4. Cross-reactivity of LacI variants toward other three untargeted inducers and IPTG. For each variant displayed in Figure 3.2, a dose-response was determined for non-target ligands and IPTG. Values displayed represent the highest fold induction at any ligand concentration. Inducers are colored as follows: gentiobiose, red; fucose, green; lactitol, blue; sucralose, magenta; and IPTG, black. Variants displayed were designed for binding to (a) gentiobiose, (b) fucose, (c) lactitol, and (d) sucralose.

Fucose- and gentiobiose-responsive variants showed pervasive cross-reactivity (Figure 3.4A and 3.4B). All four gentiobiose-responsive variants tested, including those with mutations distal to the binding pocket, showed induction by fucose as well (Figure 3.4A), suggesting that distal mutations can lead to ligand promiscuity. Variant I79Q was highly promiscuous, showing response to all five ligands (Figure 3.4B). However, mutations I79G, I79S, and I79T were specific to lactitol (Figure 3.4C), again highlighting the previously uncharacterized role of I79 in ligand-specificity determination [141-143].

Variants responsive to lactitol and sucralose were overall less promiscuous (Figure 3.4C and 3.4D); in particular the N246D/Y273H double mutant was highly specific to sucralose and showed negligible induction by any other ligand, including IPTG (Figure 3.4D). Two sucralose-responsive variants, D149G/I160V/H163Y/S193E and L148N/I160N/H163W/S191D/S193N, showed no response to fucose, lactitol or gentiobiose. The pervasive response to IPTG by nearly all engineered variants shows that response to the native inducer is robust to many mutations, and highlights the need to further improve specificity through activity maturation or to incorporate negative design against IPTG in the Rosetta protocol.

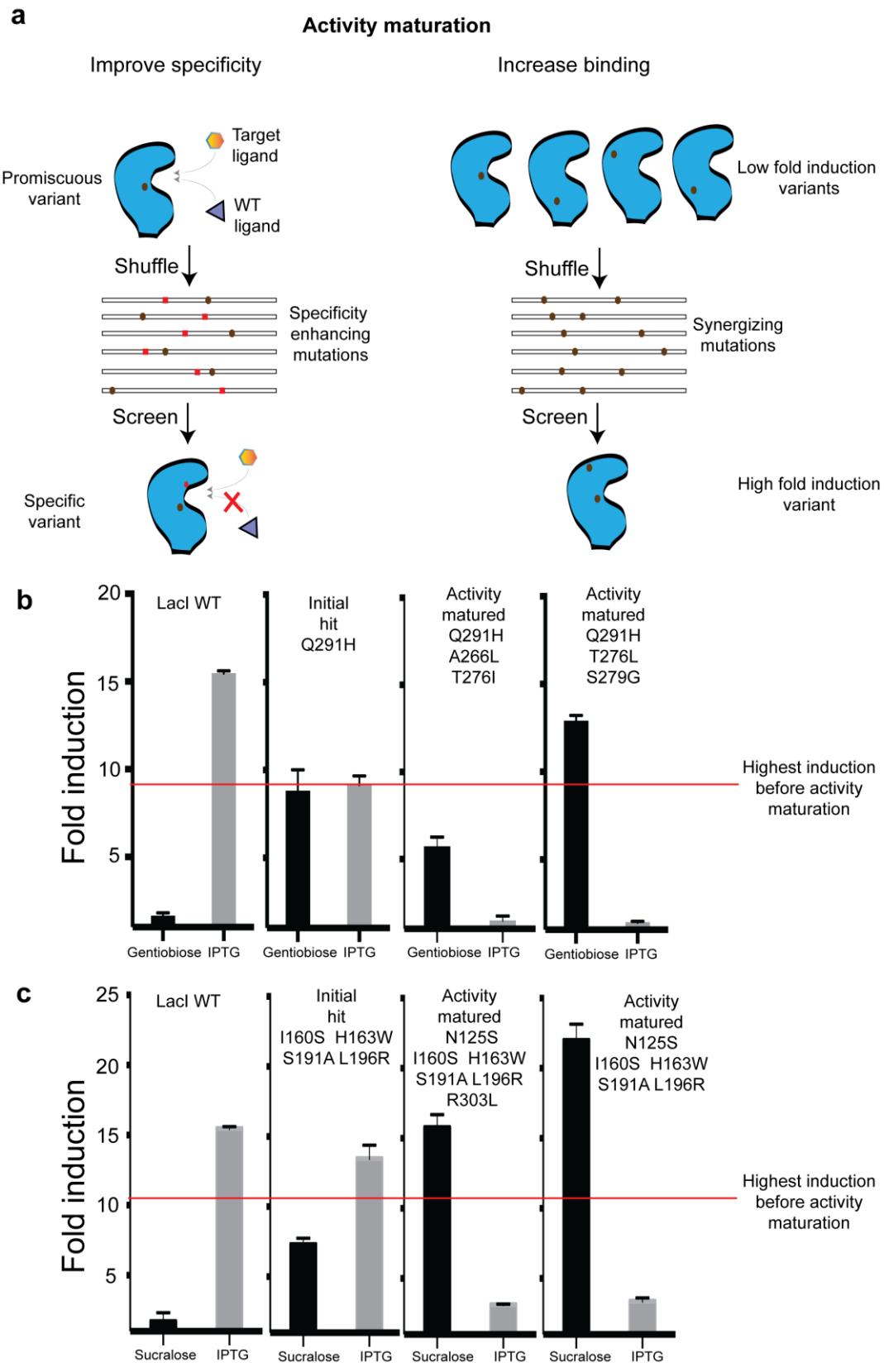
Activity maturation to improve specificity and induction

We used two approaches to mature the activity of variants: to achieve greater specificity, we individually shuffled promiscuous hits with mutations that ablate off-target binding (Figure 3.5A); and to achieve greater response, we combined multiple beneficial mutations for the same ligand (Figure 3.5A). Mutations that reduce IPTG induction, called I^S mutations [131, 132], were obtained by screening the single residue replacement library for variants that survive colicin E1 selection (bind DNA) but do not induce GFP expression in the

presence of IPTG. We sequenced 96 clones to build a library of 44 unique I^s variants with mutations near the binding pocket (Supplementary Table S3.3). We combined these I^s mutants with the gentiobiose-responsive promiscuous mutant Q291H, and screened to uncover chimeras that showed reduced IPTG induction. We uncovered variants that not only completely lost IPTG induction (Figure 3.5B), but also showed greater induction with gentiobiose.

Figure 3.5 Activity maturation of LacI variants (a) Diagram showing two activity maturation schemes for improving specificity and increasing binding affinity. (A) The first scheme (left) shuffles a promiscuous LacI variants with LacI variants that do not respond to IPTG (Is variants), and the second scheme (right) shuffles beneficial mutations from multiple variants that respond to the same ligand. The red line shows the induction level of the best hit before activity maturation (b) Induction response of WT LacI and three LacI variants toward gentiobiose and IPTG. Q291H is the promiscuous variant found after the initial screen. Activity matured variants Q291H/A266L/T276I and Q291H/T276L/S279G were found after shuffling with I^S variants. (c) Induction response of WT LacI and three LacI variants toward sucralose and IPTG. Quadruple mutant I160S/H163W/S191A/L196R is the initial hit responding to sucralose. Activity matured variants N125S/I160S/H163W/S191A/L196R/R303L and N125S/I160S/H163W/S191A/L196R were found after shuffling the initial hit with a library of other sucralose-responsive variants.

Figure 3.5 (Continued)



In another approach to activity maturation, we used gene shuffling [145] to shuffle mutations from 31 sucralose-responsive variants (Supplementary Table S3.1), and identified several clones with improved induction by sucralose (up to 22-fold; Figure 3.5C), which exceeds the LacI WT response to IPTG (15-fold). These variants also showed a dramatic decrease in IPTG induction compared to their parent sequence (3-fold versus 14-fold, respectively). The activity maturation goals of increased specificity and induction appear to be coupled, and may simultaneously improve as the binding pocket adjusts to better fit a new ligand. These results show that once an initial ligand-responsive variant is found, simple combinatorial mutational strategies can substantially improve the specificity and fold induction of the initial hit.

Crystal structure of a sucralose-binding variant

To understand the molecular details of how the LacI binding pocket accommodates sucralose, a bulky, tri-chlorinated sucrose derivative, we crystallized a computationally-designed sucralose-responsive variant in the presence and absence of sucralose (PDB IDs: 4RZS and 4RZT; Supplementary Table S3.4). This variant carries four mutations: D149T, V150A, I156L, and S193D.

In the sucralose-bound structure, the chlorines are stabilized by π-bond interactions with aromatic residues or by electron acceptor groups: glucose C3-linked chlorine is stabilized through anion-π interaction with W220, fructose C1-linked chlorine is surrounded by two π-clouds, from F161 and F293, and the fructose C6-linked chlorine is positioned around three electron-withdrawing groups: the hydroxyl group of S69 (3.5 Å); the sidechain carbonyl group of N125 (3.7 Å); and the backbone carbonyl group of N125 (3.3 Å; Figure 3.6A).

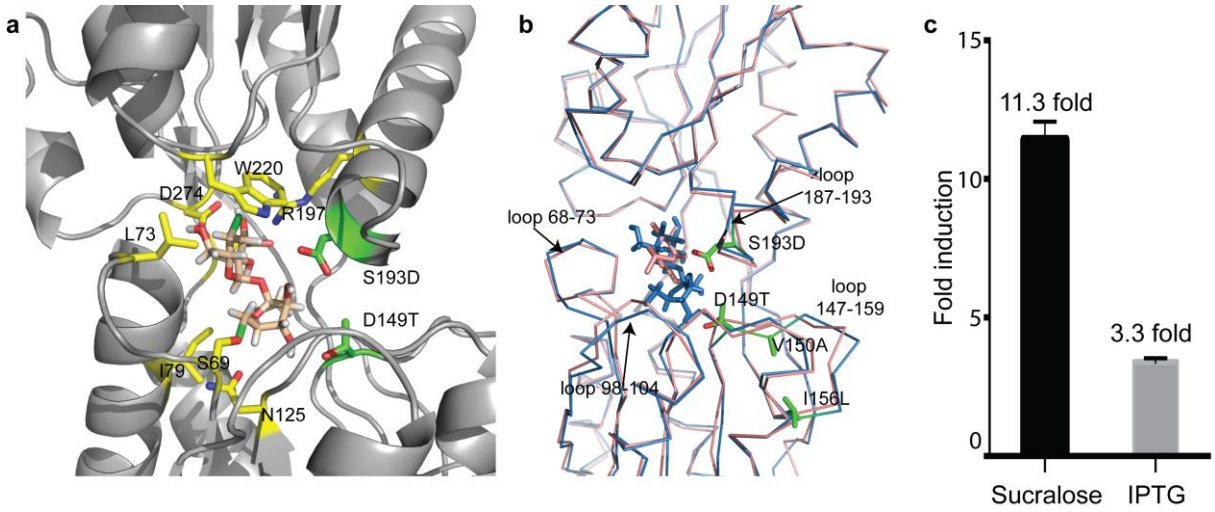


Figure 3.6. Crystal structure and GFP induction with ligand of sucralose-binding LacI design variant (D149T, S193D, V150A, I156L) (a) Close up view of sucralose bound to LacI quadruple mutant variant. Two of the designed residues, D149T and S193D, are shown in green; V150A and I156L are outside the field of view. Other key interactions of native residues are shown in yellow. (b) Backbone C-alpha structural superposition of WT LacI (pink) and sucralose-binding LacI variant (blue). The four designed residues are shown in green and loops undergoing significant conformational change are marked (c) Fold induction response of the sucralose-binding LacI design variant with sucralose and IPTG at 100 mM ligand concentration. PDB IDs of the LacI variant in apo and sucralose bound forms are 4RZS and 4RZT, respectively.

The bulkier sucralose ligand is accommodated by loops surrounding the binding pocket that are displaced outward, which affects the hydrogen bonds between the protein and the sugar hydroxyl groups (Figure 3.6B). Both D149T and S193D restore hydrogen bonds with the fructose subunit by making optimal connections and alleviating steric clash (Figure 3.6A). V150A and I156L improve sidechain packing in the loop segment 148-159, which is perturbed by sucralose binding compared to WT LacI bound to IPTG (PDB ID: 2P9H; Figure 3.6B).

This variant showed strong induction with sucralose (Figure 3.6C), but retained an un-induced baseline comparable to WT LacI, showing that allostery was not disrupted by the four

mutations. The malleability of the binding pocket to accommodate a large, chemically different inducer highlights a feature that may have evolved to allow aTFs to bind diverse ligands.

Discussion

Understanding and modifying the allosteric regulation of proteins is of considerable interest to biotechnology and medicine, given the prevalence of allostery in enzyme regulation, in the activity of protein drug targets, and in small molecule sensing. However, designing a protein to alter allosteric regulation is more challenging than designing for binding alone.

Our results suggest the following general strategy for engineering an aTF for new ligand response (Supplementary Figure S3.6). If the target ligand closely resembles a known inducer, simple mutational solutions provided by a single amino acid substitution or error-prone PCR-derived library may be sufficient. For more dissimilar target ligands, computational design is more likely to produce the more complex mutational combinations required for response. The activity of the initial hits can be further matured for greater induction or specificity as required. If no response is seen, other related aTF proteins should be tried, because the plasticity of each aTF may vary considerably. However, this aTF redesign method carries three caveats: one, it relies on the existence of an aTF crystal structure; two, the DNA-binding site for the aTF must be known; and three, the inducer for the majority of aTFs is not biochemically validated. Genomic mining can help to infer inducer identity because aTFs generally regulate biosynthetic or transport genes from neighboring loci, and each new aTF that is characterized becomes a potential starting point for biosensor design.

Besides aTFs, other biosensor approaches include riboswitches [82], reporter domain-coupled allosteric proteins [146], ligand-dependent protein dimerization [80] and ligand-conditional protein stability [69]. Despite the promise of generalizability, riboswitches have proven to be surprisingly hard to redesign or generate from aptamers. Ligand-conditional protein stability is a distinct approach to protein biosensor generation in which proteins are engineered to be stabilized through ligand binding and degraded otherwise.

Designer aTFs should find utility in many applications. Metabolic engineering approaches increasingly rely on high throughput screens and selections to identify productive combinations of mutations for metabolite production [47]. Alleviating the reliance of this approach on natural sensory proteins opens new biosynthetic opportunities for many valuable chemicals, including the identification of novel biosynthetic pathways within metagenomic libraries derived from microbes and plants. For instance, the gentiobiose-responsive LacI variant could be used to screen for β-glucosidases that carry out transglycosylation of glucose to produce gentiobiose [147]. For biocontainment applications, the survival of laboratory-engineered strains can be made synthetically dependent on sucralose, a molecule not found in nature, by placing essential genes under the regulation of a sucralose-responsive LacI variant.

New aTFs can also be powerful cell-biological discovery tools. The dynamic composition of metabolites in a cell is a signature of the phenotypic state of the cell. While the genome and transcriptome of individual cells can be interrogated, we currently lack similarly accurate tools to measure the metabolic state of single cells. Engineered aTFs that respond to key metabolites could report on the metabolic dynamics of live cells at high temporal resolution. The widely-used TetOn/Off system for mammalian gene regulation [79, 148] relies on a bacterial aTF (TetR)

adapted for mammalian cells. An expanded repertoire of similarly adapted engineered aTFs with non-interacting inducers would enable tunable, independent control of multiple genes to exquisitely regulate signaling, development and differentiation pathways. We expect the ability to create aTFs responsive to new target molecules to have wide-reaching benefits for synthetic biology.

Methods

LacI expression vector and screening strain construction

The *E. coli* strain K12 MG1655 derivative EcNR2 [30] was modified by lambda Red recombineering [126] to replace the native *lacI* gene with a zeocin resistance cassette. The *tetR* and *bla* genes found on the lambda prophage were similarly replaced with a tetracycline resistance cassette. Recombineering was then used to create the final screening strain by replacing the native promoters 5' to the *tolC* gene with a linear PCR product encoding the following: one copy of promoter pLlacO controlling transcription of a cocistron of superfolder GFP [149] and a kanamycin resistance cassette; and a second copy of pLlacO in a divergent orientation controlling transcription of *tolC*. All modifications were sequence verified.

A copy of the *E. coli lacI* gene which had been recoded to facilitate cloning site sequences (Supplementary Table S3.5) was cloned into a low-copy plasmid backbone (SC101 origin of replication) carrying a spectinomycin resistance, to create plasmid pSC101_lacI_specR (Supplementary Table S3.5). The *lacI* variant gene was expressed from the strong pLtetO promoter [45], which is unregulated in the screening strain due to the deletion of *tetR*. Colonies of the screening strain fluoresce visibly under blue light, but fluorescence was no longer visible

after transformation with plasmid pSC101_lacI_specR. Observed reversions of the repressed phenotype were low and did not necessitate restoration of the $MutS^+$ (mismatch-mediated repair competent) phenotype in this strain.

Rosetta computational design of LacI proteins

Computationally-designed LacI variant candidates were generated, using the Rosetta software suite [26, 27] for the three target ligands – fucose, lactitol and sucralose – most dissimilar to the native inducer. For each ligand, we generated a library of hundreds of allowable conformational isomers, or conformers, using OpenEye Omega software [150, 151]. These conformers sample discrete states along rotatable torsions of various bonds. Each conformer was docked into the ligand-binding pocket of a high-resolution crystal structure of LacI [144] (PDB ID: 2P9H). The design protocol consists of multiple rounds of rigid body perturbation of the ligand position followed by combinatorial mutagenesis and backbone minimization to optimize the interaction of the selected conformer within the pocket. The mutagenesis includes exhaustive sampling of the rotameric states of each amino acid in a backbone-context dependent manner. All designs were loosely filtered based on a standard set of energy terms that ensures diversity of solutions while eliminating the poor designs (see Supplementary Methods for Chapter 3 for full computational method details.)

Construction of LacI variant libraries

The *lacI* gene variant libraries were constructed by cloning oligonucleotides encoding the desired mutations into plasmid pSC101_lacI_specR amplified by PCR with primers (Supplementary Table S3.6) to appropriately linearize, add *Bsa*I recognition sequence (5'-

GGTCTCN), and remove the wild type *lacI* coding sequence to be replaced. Oligonucleotide pools [39] encoding Rosetta designed sequences were obtained from Agilent Technologies (Santa Clara, CA), and each encoded mutations within one of the following sets of *lacI* codons: 73-125, 148-197 and 245-296. Constructing the single residue replacement library involved replacing *lacI* codons from 3-359 with one missense codon encoding each of the remaining 19 natural amino acids (6802 sequences). Oligos encoding these mutations were synthesized on a B3 Synthesizer (CustomArray, Inc., Bothell, WA), and were organized into the following tiles spanning the *lacI* gene: 3-38, 39-74, 75-110, 111-146, 147-182, 183-218, 219-254, 255-290, 291-326, 327-359.

Oligonucleotides in each pool were encoded as a concatamer of: forward priming sequence; the *Bsa*I restriction site (5'-GGTCTCN); appropriate 4 base upstream overhang; encoded *lacI* mutant sequence segment; appropriate 4 base downstream overhang; the reverse complement of the *Bsa*I restriction site (5'-NGAGACC); and the reverse complement of the reverse priming sequence. Subpools were amplified using primers specific to each subpool [40] (Supplementary Table S3.6) from each oligonucleotide pool using quantitative polymerase chain reaction (SYBR qPCR master mix, KAPA Biosystems, Wilmington, MA; 20 µL reaction volume; 0.1 - 1 ng oligonucleotide pool template) until the second inflection point on a real-time plot of cycle number versus well fluorescence indicated amplification saturation was beginning, following Kosuri, et al. [40].

Error-prone PCR libraries were constructed by amplifying the *lacI* codons 67-197 with GeneMorf II polymerase (Agilent Technologies) using primers including *Bsa*I recognition sites (Supplementary Table S3.6) and 15 ng gene fragment template, which resulted in 10 µg PCR

product for 670-fold amplification and a calculated mean of 5.3 coding mutations per *lacI* gene. Subpool amplification PCR products and error-prone PCR products were digested with *Bsa*I-HF enzyme (New England Biolabs (NEB), Ipswich, MA), and appropriate plasmid backbones were digested with *Bsa*I-HF and *Dpn*I enzymes (NEB). Backbone termini were dephosphorylated with Antarctic Phosphatase enzyme (NEB). All digested nucleotides were cleaned up with Agencourt AMPure XP beads (Beckman Coulter, Indianapolis, Indiana; 1:1 ratio of beads to DNA), fragments were ligated into backbones with T4 DNA Ligase (NEB), ligation products purified with AMPure XP beads and transformed into electrocompetent *E. coli* DH10B cells (NEB). After 1 hr, a 1 µL aliquot of transformed cells was plated onto LB spectinomycin selective medium to estimate the transformed library size.

Selection and screening protocols for ligand response

For library transformations, the screening strain was made electrocompetent by harvesting early log-phase cells (10 mL per transformation at OD600 nm = 0.15-0.25), removing salt through two washes with ice cold 10% glycerol, and resuspended in 50 µL cold 10% glycerol. Ten ng of library plasmid were electroporated into the competent cells, which were recovered for 1 hr in 1 mL SOC medium. One µL recovered cells were plated on selective LB spectinomycin medium to estimate the number of transformants, and the remainder of the recovered cells were added to 10 mL of LB spectinomycin medium and selected overnight.

Screening strain cells expressing LacI variants that do not bind to operator DNA constitutively express *tolC* and *gfp*; these were eliminated through negative selection by overnight selection with colicin E1 protein. We added five µL of saturated library transformation culture to 150 µL LB spectinomycin medium supplemented with tenfold serial

dilutions of purified colicin E1 protein (2.73 mg/mL), in the range from 1:100 to 1:1,000,000; a control population of the same library was grown overnight without colicin E1. Enrichment of DNA-bound lacI variants was verified by flow cytometry next day by measuring the fraction of GFP+ cells in the colicin E1 incubated and control populations. A 1:100,000 dilution of colicin E1 (20.7 ng/mL) was generally found to be optimal.

Following negative selection, the colicin E1 selected cells were washed twice with LB and grown for 1 hr in LB spectinomycin lacking colicin E1. These cells were then subjected to a ligand response test. To carry out a response test, 1.5 uL of saturated culture was added to 150 uL of LB supplemented with spectinomycin and 3 mM concentration of the target ligand; the tested library grown in identical conditions but without ligand was used as a negative control.

Cells with GFP signal greater than the ligand-free control were collected using fluorescence-activated cell sorting (FACS) on an Avalon S3 Sorter (Propel Labs, Fort Collins, CO). Because cells expressing a LacI variant responsive to the target ligand often presented a subtle signal, we used the un-induced control to set sorting gates per library-inducer pair. After observing 100,000 cells for the induced and un-induced conditions, the sorting gate was set to maximize the difference in cells falling above the gate between the induced and un-induced conditions. This generally resulted in collecting the top 0.1 – 1% of the induced library, except where a larger proportion of cells clearly fell above the un-induced condition (e.g. for gentiobiose inductions). FACS-sorted cells were immediately recovered in LB and plated on LB-agar containing spectinomycin at several dilutions to yield a plate with hundreds of clearly separated colonies. Depending on the library, we picked 48-192 colonies into a 96-well plate to clonally test induction response. Each clone was incubated overnight with and without the

target ligand (3 mM) concentration. After 16-20 hours, the GFP response of each clone was measured with and without the ligand on a flow cytometer with high throughput sampler (LSRFortressa, Beckton-Dickenson, Franklin Lakes, NJ). The sequence of each *lacI* variant in ligand-responsive clones was determined using Sanger sequencing (Supplementary Table S3.6).

Expression and purification of sucralose-responsive LacI variant

Expression strain construction

To overexpress the sucralose-binding variant (carrying mutations D149T, V150A, I156L and S193D), we cloned into a pET14b vector (Novagen via EMD Millipore, Billerica, MA) with a constitutive T7 promoter. The lacI variant gene was cloned downstream of His and thrombin tags of the vector. We used the arabinose-inducible T7 expression host BL21-AI (Life Technologies, Inc., Frederick, MD) to avoid inducing protein expression with IPTG, which could lead to binding artifacts. We also modified the commercially available BL21-AI strain by deleting the WT copy of *lacI* gene to avoid heterodimer formation with the mutant variant. We transformed BL21-AI with the pKD46 plasmid [126], containing the lambda-Red recombineering machinery on a temperature-sensitive origin of replication (plasmid lost above 37^{o}C). We replaced the WT *lacI* gene through homologous recombination by transforming the pKD46-containing BL21-AI with a zeocin resistance cassette flanked by homology arms targeting the *lacI* endogenous locus. We induced lambda-Red expression with 1% arabinose one hour before transformation with donor cassette DNA. Following recovery, the transformed cells were plated on zeocin-containing LB-agar plates, and colonies were screened to identify WT *lacI* gene deleted strain. The pKD46 plasmid was subsequently removed by growing the cells at 37^{o}C

overnight. The pET14b vector with sucralose-binding variant was transformed into BL21-AI *lacI::zeo* for overexpression.

LacI mutant protein overexpression and purification

Several colonies of the expression host containing the sucralose-binding LacI variant were used to inoculate a 350 mL LB culture supplemented with 100 µg/mL Ampicillin and grown at 37^{o}C with 230 rpm shaking overnight to an OD600 of 4.6. 35 mL of the overnight inoculum was added to each of the six, one L Terrific broth media with 100 µg/mL ampicillin in 2.5 L shake flasks and grown at 37^{o}C at 230 rpm. The culture temperature was equilibrated to 18 ^{o}C before expression was induced using 0.5 mM IPTG when the OD600 was 2.8. The induced cultures were grown at 18^{o}C for 19 hours before they were harvested.

The cell pellet was resuspended at a 2 mL/g ratio in Buffer A (20 mM Tris-HCl, pH 8.0, 0.3 M NaCl, 10% Glycerol), supplemented with 10 mM imidazole, 2 mM βME, 2 µg/mL DNase I, 0.1 mg/mL lysozyme, 1 mM PMSF, 1 tablet/100 mL lysate cOmplete protease inhibitor cocktail (Roche, Basel, Switzerland), 5 mM MgCl. Lysis was done by sonication and the lysate was centrifuged at 35,000 x g at 4^{o}C for 30 minutes. Affinity chromatography was carried out by nutating 1 mL of HisPur Ni-NTA resin (Thermo Scientific, Waltham, MA) with the cleared lysate at 4^{o}C for 1 hour. The Ni-NTA resin was packed onto a gravity column and then washed twice by 10 column volumes (CV) of Buffer A with 10mM imidazole, once by 10 CV of Buffer A with 50mM imidazole, and then eluted twice by 10 CV of Buffer A with 0.3 M imidazole. The fractions containing predominantly target protein were pooled and concentrated using a 10 kDa MWCO Amicon Ultra-15 concentrator (EMD Millipore).

Size exclusion chromatography was subsequently performed on a HiLoad 16/60 Superdex 200 PG column (GE Healthcare, Little Chalfont, Buckinghamshire, United Kingdom) at a flow rate of 1 mL/min in Buffer A. Pure peak fractions were pooled and concentrated to 12.8 mg/mL using the same type of concentrator mentioned above. This concentrate was subject to a three step dialysis (555-fold dilution factor each step) in a buffer containing 0.2 M Tris, pH 7.4, 0.2 KCl, 1 mM EDTA, 0.3 mM DTT in 6-kDa MWCO D-Tube Dialyzer Mini dialysis devices (Novagen cat # 71504-3). The dialyzed protein solution was centrifuged at 16,000 x g at 4^{o}C for 5 minutes and was measured at 12.2 mg/mL, measured by Bradford assay.

Crystallography, X-ray data collection and structure solution

The crystal that led to the unliganded LacI variant structure was grown using hanging-drop vapor diffusion method in a 24-well VDX plate (Hampton Research, Aliso Viejo, CA). The crystal drop of 0.4 µL of the reservoir solution plus 1.6 µL 12.1 mg/mL protein concentrate was set up against 500 µL of reservoir solution, which was composed of 16% polyethylene glycol 3,350 and 200 mM ammonium nitrate. After 5 weeks of growth, the crystal grew to approximately 600 nm x 200 nm by 200 nm and was harvested by flash freezing in liquid nitrogen with 23% glycerol as cryo-protectant.

For co-crystallization experiments, sucralose was dissolved in the dialysis buffer to a concentration of 0.5 M. a small amount of the 0.5 M sucralose solution was added to the unliganded protein concentrate at a 1:150 (v/v) ratio, giving a final sucralose concentration of 3 mM and final protein concentration of 12.1 mg/mL (equivalent of 0.3 mM). Crystallization was carried out using hanging-drop vapor diffusion method. A Mosquito liquid handler (TTP LabTech, Melbourn, Hertfordshire, United Kingdom) set up a drop of 140 nL sucralose

containing protein concentrate with 70 nL of crystallization reagent against 100 µL of the same reagent in the reservoir that was composed of 0.1 M HEPES, pH 7.5, 10% polyethylene glycol 6,000, and 5% 2-methyl-2,4-pentanediol (MPD). After 4 weeks, the structure-producing crystal grew to approximately 150 nm by 150 nm by 20 nm, and was harvested with 20% glycerol as cryo-protectant.

X-ray data for both the unliganded and sucralose co-crystal structures was collected at the Advanced Photon Source (APS) beam line 24-ID-C at the Argonne National Laboratory. The data was processed using XDS [152], followed by anisotropic data removal [153]. Structure solution was found by molecular replacement using the program Phaser [150], using the core domain (residues 62 to 332) of an existing LacI structure (PDB ID: 1JYE) as the search model. Model built using the program Coot [154] and refined using the program REFMAC [155]. Translation/libration/screw [156] (TLS) vibrational motion analysis and non-crystallographic symmetry (NCS) restraints were utilized during model refinement.

LacI ortholog/paralog identification and alignment methods

We accessed a database of complete bacterial genomes from ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/ (Jan. 29, 2013) and formatted it for BLAST searching. We used BLASTP 2.2.21 [157] to search the bacterial genome database using full-length protein queries of *E. coli* LacI and full-length *E. coli* LacZ, separately, using default settings. Our LacI ortholog set contained 41 sequences from 13,591 total matches meeting the criteria: the top hit in each species, ignoring subspecies; occurred within 10 kb of any *E. coli* LacZ hit from the same species, accounting for subspecies; and had an E value of less than 0.01. The remainder of the matches were called *E. coli* LacI paralogs.

Alignments of all sequence matches from the *E. coli* LacI BLASTP query were done with CLUSTAL 2.0.12 [158] using default settings and fast pairwise alignment. From the resulting alignment, non-gap *E. coli* LacI sequence positions were used as a positional reference. Heatmap grids depicting conservation values show the percentage of utilization for the amino-acid at the *E. coli* LacI position indicated. Gaps are shown as "-" on the y-axis. Heatmap grids of induction values are the maximum weighted induction value, which scales the total induction found equally across all co-occurring mutations. Structure diagrams indicate positions with greater than two-fold induction over no ligand, colored by the frequency the position was found in the screen.

Comparison of LacI fucose-responsive mutants to GalR/S fucose-responsive proteins

To compare our designed mutations to naturally evolved aTF sequences, we relied on known fucose-responsive GalR/S proteins. We computed whether 17 LacI ligand-proximal mutations (<5 Å from IPTG in PDB ID: 2P9H) conferring greater than two-fold response to fucose were enriched among five experimentally characterized fucose-responsive GalR/S orthologs [159, 160]. After alignment, we independently calculated the frequency of amino acids at each position for 41 high-confidence natural LacI orthologs (Supplementary Figure 3.5A), and for the five GalR/S orthologs (Supplementary Figure 3.5B). We subtracted the LacI ortholog frequencies from GalR/S frequencies at every aligned position and identity, and defined differentially conserved identities to be in the top 5% of this subtracted frequency set. Within the differentially conserved identities so defined, we identified three (I79L, I79M, and 273F) of the 17 ligand-proximal fucose-responsive mutants. By Fisher's exact test, this result is significant (P = 0.0471).

Analysis of negative selection via next-generation sequencing

Single amino-acid libraries were prepared for amplicon sequencing by nested PCR amplification using a first round of PCR with primers annealing within the *lacI* gene (Supplementary Table S3.6) and a second round of PCR with primers annealing within the first set of primers and containing i5 or i7 indexing sequences and adapters for sequencing on the MiSeq instrument (Illumina, San Diego, CA) using 300 base paired-end reads. Paired sequencing reads were collapsed and filtered for sequencing errors using FLASH v. 1.2.11 with a maximum overlap of 30 [161]. Collapsed reads were then translated in three frames, aligned to wild-type protein LacI with BLAT v.35 using default settings, and the best translated alignment by percentage match was retained [162]. Protein sequences were then trimmed according to the amplicon number and known flanking sequence from the library designs. Sequences with mis-matches in the fixed flanking sequences or different than the expected length (containing insertions or deletions) were discarded. Protein sequences were then counted for pre- and post-selection, respectively, for each amplicon. Only sequences harboring a single amino-acid change found in either the pre- or post-selection were retained for further analysis. We assembled the single amino-acid sequences as rows in a table with counts for pre- and post-selection values as columns. A pseudo-count of one was added to each column. Counts between pre- and post-selection were then quantile normalized using the normalize.quantiles function from preprocessCore in R v.3.1.1. Log2 fold-changes were computed from the ratio of pre-selection divided by post-selection quantile normalized counts [163]. Each protein sequence was positioned and shown with respect to the wild-type LacI position. Final heatmap grids, associated line plot, and histogram were created in ggplot2 [164].

Gene shuffling with I^S variants for enhanced specificity

The I^S clones (nomenclature of Suckow, et al. 2006) are LacI variants that do not respond to IPTG. We sorted cells from the single amino acid mutant library that show no GFP signal when incubated with IPTG overnight. After recovery, growth and plating, we picked about 200 colonies for Sanger sequencing. This I^S set comprises variants that are allosterically broken and variants that do not recognize IPTG. Since we are interested in the latter set, we only picked variants with mutations near the binding pocket, reasoning that these are more likely to reduce IPTG binding without affecting allostery. While some of the binding pocket variants may also be allosterically broken, this set is also likely to contain variants that do not bind to IPTG. We picked about 44 variants (Supplementary Table S3.3) to form our curated I^S set. To switch specificity of the gentiobiose hit Q291H (Figure 3.5B), we amplified each of the 44 I^S variants including the backbone using PCR primers carrying Q->H mutation at 291. This gave us 44 chimeras containing Q291H and the I^S mutations. We carried out isothermal assembly of all chimeras together in a single tube. After transformation, recovery and plating, we picked about 100 clones for testing induction with IPTG and gentiobiose.

Chapter 4. Engineering allostery: a perspective

Attributions

The majority of this chapter was previously published in Trends in Genetics [130], and does not require explicit permission to reproduce here.

Srivatsan Raman, Noah Taylor and Stanley Fields and George M. Church conceived the article. Srivatsan Raman, Noah Taylor, Naomi Genuth and Stanley Fields wrote the paper.

Summary

Allosteric proteins have great potential in synthetic biology, but our limited understanding of the molecular underpinnings of allostery has hindered the development of designer molecules, including transcription factors with new DNA-binding or ligand-binding specificities that respond appropriately to inducers. Such allosteric proteins could function as novel switches in complex circuits, metabolite sensors, or as orthogonal regulators for independent, inducible control of multiple genes. Advances in DNA synthesis and next-generation sequencing technologies have enabled the assessment of millions of mutants in a single experiment, providing new opportunities to study allostery. Using the classic LacI protein as an example, we describe a genetic selection system using a bidirectional reporter to capture mutants in both allosteric states, allowing the positions most crucial for allostery to be identified. This approach is not limited to bacterial transcription factors, and could reveal new mechanistic insights and facilitate engineering of other major classes of allosteric proteins such as nuclear receptors, two-component systems, G protein-coupled receptors, and protein kinases.

Unlocking the power of allostery in synthetic biology

Allosteric regulation mediates virtually every biological process, including transcription, signal transduction, and enzyme activity and transport. Allostery can be broadly defined as activity at one site in a protein regulating function at a spatially distant site. Allosteric regulation occurs through an allosteric effector, generally a small molecule, which binds at one active site and triggers a conformational change that affects function at a distant site. Because of their ability to respond to small molecules by a change of state, allosteric proteins play an important role in synthetic biology. Nevertheless, our ability to engineer allosteric proteins is highly constrained by our limited understanding of the molecular details of allostery, and thus we have barely scratched the surface of how allosteric proteins can be applied in this emerging field.

Allosteric proteins are used as switches in synthetic circuits. Although synthetic biologists would like to build more complex circuits, a major limitation is the lack of orthogonal switches (allosteric proteins that bind to different inducers and different DNA sequences with little crosstalk). A suite of well-characterized orthogonal switches would vastly enhance our ability to build higher-order synthetic circuits with real-world applicability [127]. For example, such switches could serve as analog-to-digital converters that convert a continuous chemical gradient into a digital output. Bacteria possessing synthetic circuits combining many such analog-to-digital converters could then be used as whole-cell biosensors of the gut [127].

Allosteric proteins can also be used as in vivo metabolite sensors for engineering biosynthetic pathways [47]. These sensors detect and respond to the level of some sought-after metabolite, enabling genetic selections in which the best producers are identified from a large

number of variant organisms. Despite an increasing demand for allosteric sensors detecting industrially useful chemicals, we are limited to the ligand-binding domains of known transcription factors; however, this bottleneck could be removed by designing new allosteric proteins. For instance, new small-molecule sensors could be generated from chimeras of a well-characterized DNA-binding domain with ligand-binding domains identified in the sequences of metagenomic samples. Alternatively, we might be able to mutate binding-site residues in an existing sensor to create new ligand specificities without affecting allosteric communication [165].

Besides their biotechnological applications, designer allosteric proteins can provide independent, temporal regulation of multiple genes, a useful tool for developmental biology. The Tet-on/off activator system, based on the *Escherichia coli* tetracycline resistance (Tet) repressor, is widely used for mammalian gene regulation [79, 148], but it does not allow the independent control of multiple genes. With multiple orthogonal regulators similar to the Tet-on/off elements we could, for instance, gain exquisite control over stem cell differentiation pathways by modulating each differentiation factor independently.

Finally, redesigning allosteric proteins to respond to molecules that cross the blood–brain barrier would enable the activation of specific neural circuits in the brains of live animals simply by incorporating the inducers in the diet. To engineer allosteric proteins, however, we need to take a closer look at how allostery works at the molecular level.

Efforts to understand allostery have largely focused on biophysical models to explain the conformational transition between two states, corresponding to the presence and absence of the effector [4]. Protein dynamics shows that allosteric transitions occur as a consequence of

local conformational changes such as disorder or unfolding that are propagated to distal regions, shifting the overall conformational equilibrium [127, 166-168]. Whereas these models focus on the thermodynamic drivers of allostery, little is understood about the underlying molecular basis. Alternatively, a genetic approach to understanding allostery could be taken, by making large numbers of single mutations and combinations of double mutations in an allosteric protein, and then determining which mutant proteins maintain allosteric coupling. This approach, in conjunction with biophysical measurements derived from nuclear magnetic resonance (NMR) or molecular dynamics simulations, may tell us how proteins propagate allosteric control, and might let us develop rules for engineering allosteric proteins to use in synthetic biology applications. The grand goal would be to engineer de novo an allosteric transition in a protein that is not normally allosteric. For instance, could we convert an immunoglobulin domain to actuate an allosteric response when it binds to an antigen? While this may appear a daunting challenge, a near-term achievable goal might combine domains from two different allosteric proteins to result in a functional chimera (Figure 4.1).

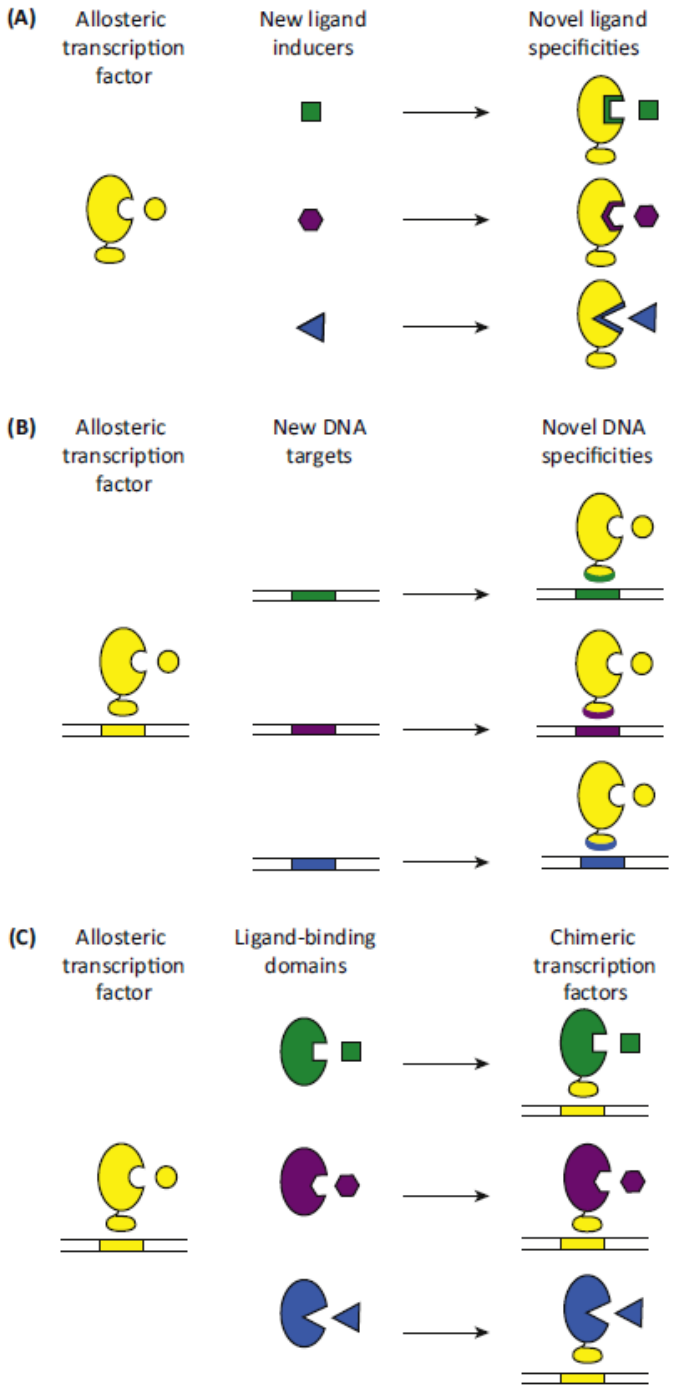


Figure 4.1. Engineering novel allosteric proteins. (A) Designing novel ligand specificities: a known allosteric protein can be made to recognize different small-molecule inducers by mutating the amino acids in the ligand-binding domain. (B) Designing novel DNA specificities: allosteric transcription factors can be targeted to new DNA sequences by altering the DNA-binding domain. (C) Novel allosteric chimeras: allosteric protein domains that are not capable of binding to DNA, such as periplasmic-binding proteins, can be attached to a DNA-binding domain to create novel, chimeric transcription factors.

Our ability to engineer proteins to bind to DNA, small molecules, or other proteins draws from a wealth of research studying these interactions at molecular resolution. However, allostery has proved to be recalcitrant to engineering because we do not fully understand the molecular connectivity involved in allosteric communication. Previous studies indicate that a network of structurally contiguous residues act in concert to transmit the allosteric signal [127, 169-172]. We review here a classic allosteric regulator and describe how new technologies – based on protein-wide mutational perturbation – could build a molecular 'wiring diagram' of allostery by deconstructing the role of each amino acid in the transmission of the allosteric signal. Knowledge of the wiring diagram should allow us to preserve allosteric connections as we design new functions into an allosteric protein. We conclude with how similar strategies might be applied to other broad classes of allosteric proteins.

Lessons from LacI

Allosteric transcription factors in bacteria, one of the largest annotated families of proteins, regulate adaptive responses to environmental cues. The best- and longest-studied allosteric protein is the *E. coli* repressor LacI which regulates the *lac* (lactose-utilization) operon [144, 173]. LacI is composed of ligand-binding and DNA-binding domains. In the absence of the ligand, LacI has high affinity for DNA; when bound to inducer, the protein undergoes a conformational change that causes the DNA-binding domain to lose affinity for DNA, thereby dissociating from the operator site and unblocking the path of RNA polymerase to transcribe downstream genes.

The structure of LacI [144, 173] can be divided into three sections: the N-terminal 60 residues form a helix-turn-helix motif that binds to DNA; the core of the protein (residues 61–

330) is made up of N- and C-terminal subdomains where the ligand binds; and the C-terminal 30 residues are involved in tetramerization (Figure 4.2). LacI is a functional dimer that makes extensive monomer–monomer contacts across the N- and C-terminal ligand-binding subdomains [144, 173]. The ligand binds in the cleft between the core subdomains, inducing a Venus flytrap-like allosteric motion. Upon induction, the allosteric signal is communicated by the relative motion between the N- and C-terminal ligand-binding subdomains, causing the DNA-binding domain to undergo a helix-to-coil structural transition, and hence lose its DNA-binding activity.

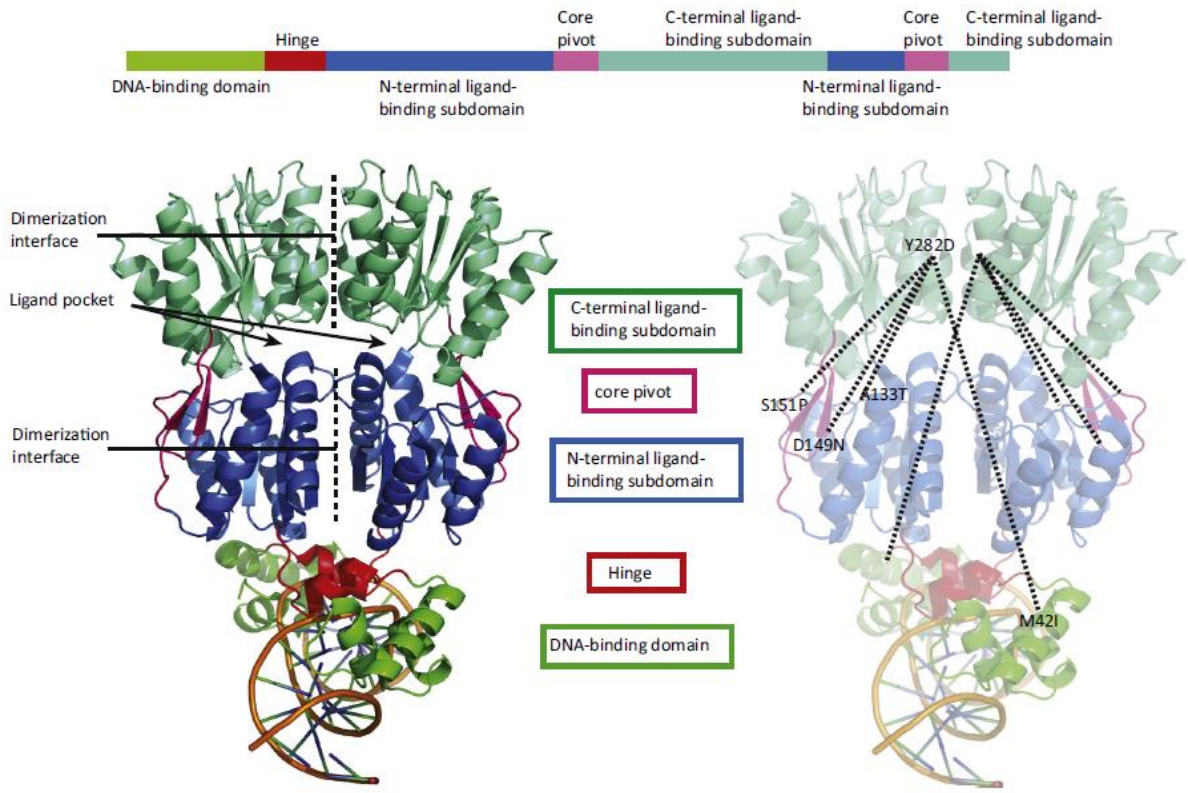


Figure 4.2. LacI protein architecture and long-range allosteric connections. Structure of a LacI dimer bound to DNA is shown on the left side; key regions are highlighted in different colors. The broken lines on the right side show long-range connection between four mutations (M42, A133, D149, and S151) that rescue allostery in Y282D.

The molecular basis of allosteric communication in LacI has been the focus of extensive genetic and biochemical studies. Jeffrey Miller's laboratory made 4000 point mutants of LacI and used a sensitive genetic screen to classify each mutant into one of three phenotypes: wild type-like activity (WT); unable to bind to DNA (I^-); or unable to bind to inducer or broken allostery (I^S) [131, 132, 174]. These studies showed how structural integrity, dimerization, allosteric signal transduction, ligand-binding, and DNA-binding are tightly interlinked and act in concert for protein function.

The I^- phenotype was generally caused by mutations in the DNA-binding domain, or in the C-terminal ligand-binding subdomain which acts as a scaffold for the dimerization required for DNA-binding. Long-range allosteric communication was evident when over 20 second-site mutations scattered across the structure reverted the I^- phenotype of Y282D, a change that disrupted dimerization [139, 140] (Figure 4.2). These second-site mutations were clustered in the core-pivot region.

The I^S phenotype resulted from mutations that affected the extensive interactions between the DNA-binding domain and the N-terminal ligand-binding subdomain. Mutation at a single position, A110, in the dimerization interface of N-terminal ligand-binding subdomain dimerization interface caused opposite phenotypes depending on the substitution: higher inducer affinity and lower DNA affinity, or lower inducer affinity and higher DNA affinity, highlighting the key role of the interface in allostery [175].

Co-evolution is a molecular signature of functionally coupled residues undergoing coordinated sequence changes across multiple members of a protein family. To preserve protein function, co-evolving pairs undergo mutually compensatory changes during evolution,

which provide insight into residue connectivity [176]. The LacI family of bacterial transcription factors, despite sharing the same protein architecture, have a sequence similarity of less than 30% [177]. Although residues in the ligand-binding pocket and at the DNA-binding interface vary depending on inducer and operator sequence, residues involved in allostery have lower sequence entropy. Indeed, most co-evolving residues involved in allostery within a subfamily were found in spatially disconnected regions, showing that complex epistatic networks participate in allostery [130, 178]. In LacI, these positions are scattered across the N- and C-terminal dimerization interfaces, core-pivot and hinge regions, and DNA-binding domain, consistent with genetic and biochemical studies of the mutants [177].

Within a close subfamily, structural residues in the ligand-binding pocket are more conserved and ligand specificity is achieved through mutations at less-conserved residues. Comparison across more distant families shows that approximately 19% of the residues are conserved, suggesting that these positions may be indispensable or 'hardwired' for structural integrity and function for that protein-fold family [177]. As the scaffold evolved to acquire specific functions, it may have adapted to a particular niche through minor alterations in mechanism. For instance, although LacI and PurR share highly similar structures and sequences, their allosteric wiring is opposite: upon ligand-binding, LacI dislodges from DNA, whereas PurR binds to DNA.

Deep mutational scanning – a genetic approach to understanding allostery

Biochemical and evolutionary studies of LacI suggest that amino acids involved in allostery are intricately coupled with amino acids recognizing ligand and DNA. Thus, a first step toward unraveling this allosteric network would be to determine the role of all the amino acids

in LacI function, classifying each as participating in one of the following categories: binds to DNA or to ligand, is required for structural stability, is involved in allostery, or none of the above. As with conventional protein biochemistry, the functional importance of each amino acid can be determined by mutational analysis. Because allostery is a systemic effect, however, we need to scale-up mutagenesis by orders of magnitude to interrogate the effect of every single mutation and every combination of double mutations. The double mutations are particularly important because they can reveal long-range allosteric interactions, as seen with the Y282D mutation in LacI [139, 140] . Although allostery most likely depends on higher-order network interactions beyond pairwise couplings, sufficient pairs of long-range double mutants should identify linchpin positions in this network.

How can we obtain the phenotype associated with every single and double mutation in an allosteric protein? An approach called 'deep mutational scanning' [179, 180] can generate a comprehensive mutation map which shows the effect of all neutral and loss-of-function, and any gain-of-function or hyper-activating changes, as well as double mutations that restore function lost in a single mutant or that exacerbate the combined effect of single mutations more than predicted. Deep mutational scanning has successfully been used to analyze antibody affinity maturation [181, 182], protein–peptide interactions [183, 184], protein–small-molecule affinity [92], ubiquitination [185], protein stability [179], splicing [186], and antibiotic resistance [187], among other protein activities.

Deep mutational scanning leverages two major advances: first, the generation of targeted DNA libraries with over a million unique sequences through doped oligonucleotide assembly or pre-specified microarray-generated oligonucleotides [40]; and, second, enormous

read numbers from next-generation sequencing that allow thousands of clones to be assayed simultaneously by linking activity to read-frequency of each unique sequence [179]. Genotype and phenotype are linked through an *in vitro* or *in vivo* selection that amplifies mutants with the desired function. Because the total sequence read-count is orders of magnitude greater (e.g., 400 million reads with the Illumina HiSeq) than the number of selected genotypes, the activity of a protein variant, to a first approximation, is reflected in the relative sequence abundance of its encoding DNA. Enthalpies derived from sequencing statistics have been shown to linearly correlate with experimental measurements and computational structure-based Gibbs free energy change (ΔG) predictions [182, 188]. For LacI, the library of all single mutations is a little over 6000, and even the library of all double mutations, at nearly 2.5×10^7, is within the transformation efficiency (10^9) of commercially competent cells.

Once the mutants are constructed, they need to be assayed in a suitable selection system. Selection systems typically enrich for only one function; for instance, with protein–peptide interactions, phage display enriches for stronger binders. However, a selection system for allosteric proteins should enrich for either of two states: the induced or the un-induced. Such selections can be accomplished using a dual selectable marker as the reporter gene. For example, the *E. coli tolC* gene [31, 111] and the *Saccharomyces cerevisiae URA3* gene allow enrichment of cells expressing (positive selection) or not expressing (negative selection) the reporter using different selection agents. Dual selection can thus identify mutants with the three phenotypes found in Miller's study: WT, I^-, and I^s [132].

First, subjecting the library to positive selection in the absence of the inducer reveals mutants that are not capable of binding to DNA (I^- phenotype) because they constitutively

express the reporter and are enriched (Figure 4.3, left arrow). Second, subjecting the library to negative selection in the absence of the inducer identifies mutants that bind to DNA; these repress the reporter and are insensitive to negative selection (Figure 4.3, center arrow). Subjecting this DNA-bound population to positive selection in the presence of the inducer enriches for mutants with WT phenotype because they respond to the inducer to activate the reporter (Figure 4.3, center arrow). Finally, exposing the DNA-bound population to negative selection in the presence of the inducer enriches for mutants that do not respond to the inducer (I^s phenotype), while those exhibiting WT-like activation perish (Figure 4.3, right arrow).

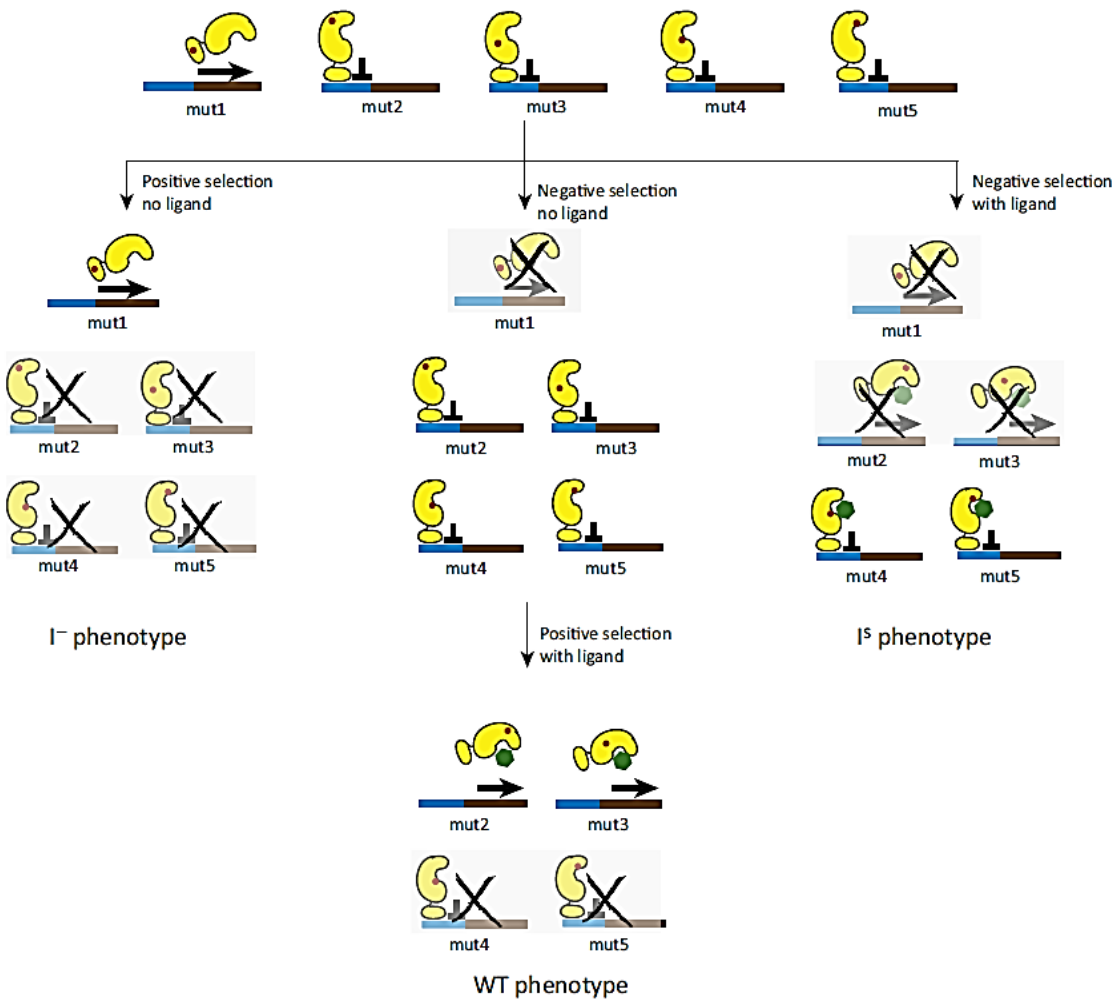


Figure 4.3 Toggled selection scheme for high-throughput functional evaluation of LacI mutants. LacI mutants ('mut') are shown in yellow; red circles represent mutations. Positive selection in the absence of the inducer enriches for I^- mutants (left arrow). Negative selection without inducer followed by positive selection with inducer enriches for wild type (WT)-like mutants (center arrow). Negative selection with inducer enriches for I^S mutants (right arrow).

The next step is to identify specific positions from the I^- and I^S set that play a role in allostery. To delineate this subset, we need to reclassify I^- and I^S into subgroups. The I^- set comprises mutants that do not recognize DNA, structurally unstable mutants, and mutants that have broken allostery. The I^S set comprises mutants that do not recognize the inducer and those that fail to transmit the allosteric signal despite inducer binding. To narrow-down the

allosteric subset from the I^- and I^s sets, we can carry out computational structure-based ΔG calculations of folding, ligand-binding, and DNA-binding of each mutant. Because the allosteric subset is deduced by the process of elimination, we need to prune this set further to isolate a minimal subset of positions that maximally contribute to allostery. NMR chemical shift perturbation measurement detects the local change in chemical environment for each amino acid upon conformational change. This approach has been used to identify the allosteric connectivity in protein kinase A [172]. Because amino acids in the vicinity of the allosteric positions are likely to undergo conformational change, chemical shift perturbations alone cannot identify the allosteric network. However, via the intersection of mutational data from genetic screens, structure-based enthalpy calculations, and biophysical measurements of local conformational change by NMR, we can pinpoint the minimal set of amino acids that constitute the allosteric network. Molecular simulations can independently validate the allosteric network by computing the conformational change of individual amino acids across the two allosteric states [169, 189].

Thus, in one multiplexed assay, millions of LacI variants can be functionally characterized in a manner that is facile and might be extendable to other proteins. Extending this selection system for other members of the LacI family might be accomplished by swapping the promoter regulating the reporter with a promoter that is bound by a different allosteric protein.

Engineering allosteric proteins

New ligand and DNA specificities engineered into existing proteins

To engineer an orthogonal transcription factor, we can begin by redesigning the ligand- and DNA-binding specificities of natural transcription factors. Once the allosteric connections have been identified, and the residues responsible for allostery have been distinguished from those involved in binding ligand or DNA, we can incorporate this information into the design protocol. For example, in the case of LacI, mutating residues that cause an I^s phenotype is more likely to result in altered specificity because a subset of I^s mutants retain allostery despite losing binding affinity for the inducer, isopropyl β-D-1-thiogalactopyranoside (IPTG), a lactose analog. By contrast, to change the DNA specificity, we target I^- mutants for additional mutations that restore allostery and to achieve new specificity because both are intimately linked in the DNA-binding domain of LacI.

Despite different ligand specificities, the overall binding-pocket architectures of LacI-like proteins are similar. This similarity suggests that an altered specificity for a new ligand can be achieved while still preserving allostery. The binding pocket of each allosteric protein family can, in principle, be redesigned to access chemical diversity around its cognate ligand. For instance, the LacI family of proteins may be able to accommodate many derivatives of sugars. With respect to DNA recognition, the helix-turn-helix domain is commonly used across several bacterial transcription factor families. Thus, by analyzing the binding-site preferences across helix-turn-helix family members, we might gain the knowledge to engineer new DNA specificity without compromising allostery.

Chimeric allosteric proteins

An alternative to redesigning existing proteins is to engineer new allosteric proteins by mixing and matching protein domains. The LacI family is thought to have evolutionarily diverged from the structurally similar periplasmic-binding protein family by acquiring a DNA-binding domain [190]. Periplasmic-binding proteins respond to a large repertoire of small molecules by regulating ABC transporters, and can be engineered for new bonding using computational methods [65]. A chimera of a periplasmic-binding protein and a DNA-binding domain could be engineered such that an allosteric change in the periplasmic-binding protein is communicated to the DNA-binding domain to change DNA affinity. Another DNA-binding domain family that might be amenable to engineering is the zinc-finger protein class, which is comparable in size to the helix-turn-helix domain class. Because the specificity of these proteins can be programmed [191], we can build completely orthogonal switches. A further approach is to engineer a chimeric protein that, upon ligand binding, exposes a binding surface that recruits a sigma factor, resulting in the activation of transcription. Engineering functional chimeras will involve extensive optimization of the linker between the domains, the inter-domain interface, and the dimerization interface.

Although thousands of bacterial allosteric transcription factors have been annotated in the sequences of metagenomic samples [90, 91], their ligand and DNA specificities remain unknown. However, a clue to the type of ligand that is recognized might come from characterization of the nearby operon that is regulated by a transcription factor. This set of transcription factors is a treasure trove for the design of biosensors for many industrially valuable molecules, allowing their biosynthetic pathways to be optimized for higher production.

If the rules of allostery are known, we can build a functional chimera with a known DNA-binding domain and identify its inducer using a reporter-gene assay against a panel of ligands.

In general, knowledge of allosteric connections can help favorably bias the search for new function toward retaining residues that mediate communication. Computational and directed evolution design protocols could incorporate deep mutational scanning data when choosing residues that can be mutated.

Application to the wider world of allostery

As DNA synthesis and sequencing become increasingly cheaper and higher in throughput, the rate-limiting step for the analysis of other classes of allosteric proteins is the development of functional assays. The bacterial transcription factor family allows one of the easiest functional assays because allostery is directly coupled to transcription. In this section we briefly describe functional assays for other major classes of allosteric proteins.

Direct transcriptional readout (nuclear receptors)

The toggled selection scheme for LacI can be adapted for other allosteric transcription factors such as nuclear receptors (Figure 4.4A). Binding of the ligand to the nuclear receptor induces a reorientation of the ligand-binding domain and, in the case of steroid receptors, directly leads to transcriptional activation. Other nuclear receptors, such as the receptors for retinoic acid and vitamin D, may also be successfully interrogated with this strategy, although upon ligand binding they have a more complex set of binding interactions involving additional transcriptional activators and repressors [192-194].

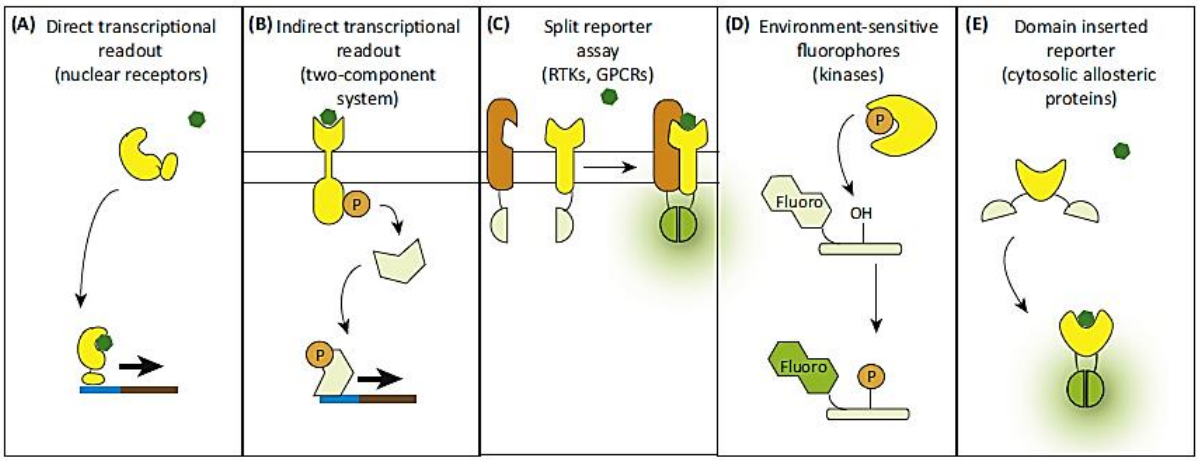


Figure 4.4. Functional assays for other allosteric protein classes. (A) Direct transcriptional readout: applicable to allosteric transcription factors such as nuclear receptors. The presence of the receptor ligand is necessary to activate expression of a selectable marker. (B) Indirect transcriptional readout: applicable to two-component systems. Allosteric activation of a histidine kinase causes phosphorylation (P) of a response regulator, leading to transcription of a selectable marker. (C) Split reporter assay: applicable to G protein-coupled receptors (GPCRs) and heterodimeric receptor tyrosine kinases (RTKs). Complementary halves of GFP are attached to the receptor and to the factor recruited upon allosteric activation. A functional GFP is formed only when the activated allosteric protein binds to its partner and brings the two GFP halves into sufficient proximity. (D) Environment-sensitive fluorophores: applicable to protein kinases. A fluorophore is attached to a peptide substrate of the kinase in close proximity to the phosphorylated residue, such that the fluorophore will fluoresce only when the phosphate group is present. (E) Domain-inserted reporter: applicable to cytosolic allosteric proteins. The allosteric protein is inserted into GFP such that a functional GFP is formed only when the allosteric protein changes to a ligand-bound conformation.

Indirect transcriptional readout (two-component system)

Many allosteric proteins alter gene expression indirectly by regulating downstream transcription factors. A transcriptional readout may be employed to evaluate such allosteric proteins provided that target gene expression robustly reflects protein activity (Figure 4.4B). An example of such pathways is the bacterial two-component system containing a membrane-bound histidine kinase, that serves as an environmental sensor, and a cytoplasmic response regulator, which is phosphorylated by the activated kinase and often acts directly as a

transcription factor [195]. Systematic mutagenesis could identify the amino acids needed for the sensor histidine kinase to change conformation upon ligand activation [88, 196] by employing a toggled selection marker at the locus that is bound by the response regulator as a readout of pathway activation.

Split reporter assay (GPCRs and RTKs)

For allosteric receptors far upstream of the transcription factors in their pathway, a different readout is needed. Membrane-bound receptors such as G-protein coupled receptors (GPCRs) and receptor tyrosine kinases (RTKs), upon allosteric activation, colocalize with a partner: GPCRs associate with β-arrestin [197] and EGFR family RTKs can form heterodimers [198]. Activity of these proteins can be measured with a split reporter assay in which the receptor is fused to one domain of a reporter protein and the recruited partner is fused to the other domain; when brought into close proximity the two domains can combine into a functional whole (Figure 4.4C) [199-202]. This assay can be performed using a divided GFP and fluorescence-activated cell sorting (FACS) to sort cells based on fluorescence intensity.

Environment-sensitive fluorophores (protein kinases)

Allosteric protein kinases can be difficult to study owing to the transitory nature of their interactions with their substrates. Kinases that do not directly regulate transcription factors may be interrogated using fluorescent sensors dependent on target phosphorylation. For tyrosine kinases whose phosphorylated targets are recognized by SH2 (Src homology 2) domains [203], the kinase target can be fused to SH2, together with two fluorescence resonance energy transfer (FRET)-compatible fluorophores at either end of the construct, such

that SH2 binding to the phosphotyrosine brings the fluorophores together. Alternatively, as a more general tyrosine/serine phosphorylation sensor, a peptide substrate of the kinase is attached to a fluorophore with a separation of only several angstroms; when the peptide is phosphorylated, the change in the microenvironment activates the fluorophore (Figure 4.4D) [87, 204, 205]. FACS can then be used to sort cells based on fluorescence intensity.

Domain-inserted reporter

As a general methodology for cytoplasmic allosteric proteins, various domain-inserted reporter systems have been developed. The allosteric protein is inserted into a reporter protein in such a way that reporter activity is dependent on the conformation of the allosteric protein (Figure 4.4E). β-Lactamase has been used as such a reporter with maltose-binding protein, leading to maltose-inducible ampicillin resistance [146]; similarly, a calcium-sensitive GFP reporter was made through insertion of calmodulin [206]. In this way, a wide array of allosteric proteins, such as enzymes and ion sensors that do not regulate transcription, can be tied directly to fluorescent and antibiotic-resistance readouts.

Concluding remarks

The ability to engineer the above allosteric protein classes paves the way for new synthetic biology applications: designer GPCRs that can respond to a drug overdose, two-component system proteins that enable bacterial chemotaxis toward a specific molecule, dynamic rewiring of kinase signaling, and controlling the composition and function of engineered microbiota with quorum-sensing switches.

The approach outlined here shows how the power of deep sequencing can be harnessed to address a longstanding question in biology: how protein sequence affects allostery. We envisage that this rich mutational dataset will motivate new studies in kinetics of allostery through molecular dynamics and NMR experiments.

Chapter 5. Conclusion

Treating biological engineering like traditional, 'harder' engineering disciplines, and assuming that systems will behave well enough to design them bottom up, is an attractive paradigm that has pushed the field to probe the rules that govern basic biological processes like transcription [7], translation [6], protein expression [207], and protein folding [208]. Enzymes can be designed *in silico* that behave as expected once expressed in a cell [26], an achievement that was unthinkable just two decades ago. Metabolic design algorithms [11], ever more complicated genetic circuits [127], and even dedicated programming languages [209] have been devised to tame the dizzying complexity of cellular machinery in order to design and harness its power for human needs. These successes validate this approach, and promise a bright future for bottom-up biological engineering.

But in many ways, biology is very different from other systems where bottom-up engineering has succeeded for decades. For the foreseeable future, the most robust (and perhaps only) living systems will remain those that are based off of existing cells. This means that most of the components were likely repurposed from some earlier function which they may partially or wholly retain. Each genetically-encoded protein or RNA can interact with hundreds of others in cryptic, competitive or synergistic ways. The arising complexity and redundancy explains much of the resilience of cells and organisms to changing environments: redundant pathways are robust to mutations that may disrupt function, and complex interactions can help new functions evolve in few generations. So perhaps it is futile to abstract biological engineering away from biological complexity.

Instead, we can harness evolution as a biological engineering paradigm to coax a population of cells to develop a function we desire. Directed evolution is indeed one of the oldest forms of biological 'engineering,' but it is only recently gaining traction for application to whole biosynthetic pathways as DNA construction, high-throughput screening (e.g. FACS) and whole genome sequencing have caught up with metabolic design. With the right incentive, cell populations can be directed to evolve chemical tolerance [37, 38] or metabolite production phenotypes. Synthetic auxotrophies, like those we demonstrate in Chapter 2 [43], provide such an incentive, allowing metabolic engineering to proceed in a manner much more agnostic to the function and interaction of each protein in the cell. All of the knowledge useful for designing bottom-up biological systems is still useful when directing population evolution from the top down, and greatly speeds up the rate of success, but, it is not so essential that each part be understood well enough to be computationally modeled.

As discussed, biosensor-directed metabolic engineering has shown the capability to improve a number of biosynthetic pathways. So far, though, the gains have been modest, from a few fold improvement in production titer [47, 66] to a few tens of fold improvement [43]. This is still a far cry from the thousands of fold improvement from initial pathway demonstration to commercial strain production that have been realized through more traditional metabolic engineering approaches [4]. To compete, sensor-directed pathway engineering will require a much more ambitious approach to whole pathway modification and dynamic control, enabled by a wealth of new biosensors toward pathway intermediates and central metabolites. These are daunting challenges, but the power of multiplexed screening of billions of metabolic designs is begging to be explored.

Appendix

Supplementary methods for Chapter 3

Computational protein design protocol with Rosetta

Local Rosetta runs were performed with a developers' version of Rosetta, corresponding to a version slightly after the Rosetta 3.2.1 release. Rosetta@Home runs were performed in July and August of 2011, using Rosetta@Home version 3.14.

Protein Preparation

The highest resolution structure of *E. coli* LacI co-crystallized with IPTG was selected (PDB ID 2P9H) [210] This structure has two protein chains in the asymmetric unit - chain A was arbitrarily chosen to be the template structure on which to do the designs.

The template structure was idealized and relaxed in the presence of the native IPTG ligand using an early version of the all atom coordinate constrained relax protocol [211], but with more restrictive restraints.

Commands:

```
./sidechain_cst.py 2p9hA.pdb 0.25 0.35;
~/rosetta/rosetta_source/bin/idealize.linuxgccrelease \database
~/rosetta/rosetta_database/ -in:file:fullatom -s 2p9hA.pdb \
extra_res_fa 2p9hA_lig.params -no_optH false -flip_HNQ;
~/rosetta/rosetta_source/bin/relax.linuxgccrelease -database \
~/rosetta/rosetta_database/ -relax:sequence_file \
always_constrained_relax_script -
constrain_relax_to_native_coords \ relax:coord_cst_width 0.25 -
relax:coord_cst_stdev 0.25 \  constraints:cst_fa_file
2p9hA_sc.cst -s 2p9hA_idealized.pdb \  in:file:native 2p9hA.pdb
-extra_res_fa 2p9hA_lig.params \  in:file:fullatom -no_optH
false -flip_HNQ;
```

Ligand preparation

3D structures of the design ligands (fucose, arbutin, lactitol and sucralose) were created with Spartan (Wavefunction, Inc.), and heavy atom conformational libraries were made with OpenEye Omega [151]. Fucose was modeled separately in both the alpha and beta conformations, with a single heavy atom conformer each. Lactitol was modeled with a generous energy window.

Commands:

```
omega2 -in ${n}.mol2 -out ${n}_confs.mol2 -commentEnergy \
includeInput -enumRing false -ewindow 70 -maxconfs 1000;
```

Ring sampling was allowed for sucralose to permit conformational changes in the five member ring. Conformations were post-filtered to remove those which altered the six-member ring conformation.

Commands:

```
omega2 -in ${n}.mol2 -out ${n}_confs.mol2 -commentEnergy \
includeInput -ewindow 30 -maxconfs 1000 -verbose -rms 0.3;
```

Rosetta topology files were generated using the molfile_to_params.py application on the heavy atom conformers, using AM1-BCC partial charges. (https://www.rosettacommons.org/manuals/archive/rosetta3.2_user_guide/app_ligand_dock.html#prep_ligand)

To ensure that Rosetta appropriately overlays the six-member sugar ring when substituting conformers, an "mctrl" file was used to specify the root and "neighbor" atom to be the anomeric carbon, if the autoselection algorithm would have chosen a non-ring atom.

```
Commands
~/rosetta/rosetta_source/src/apps/public/ligand_docking/assign_c
harges.py < ${n}_confs.mol2 > ${n}_charges.mol2;

~/rosetta/rosetta_source/src/python/apps/public/molfile_to_param
s.py -n LIG -k ${n}.kin -p ${n} ${n}_charges.mol2 --m-ctrl
mctrl;
```

The individual ligand conformer structures were overlaid on the six-member sugar ring of IPTG using the obfit program of OpenBabel [212].

Commands:

```
for f in ${n}_????.pdb
do
obfit C1CCCCO1 ${reference}.pdb ${f} > fit_${f};
done
```

These conformers were filtered for backbone clashes by examining the Rosetta-calculated fa_rep value between the ligand and an all-alanine mutant of LacI. (Produced by using a text editor to change all non-glycine residue names in the LacI structure to "ALA" - when read in by Rosetta, extra atoms will be ignored.) Conformers with fa_rep values greater than 1.0 were discarded.

Commands:

```
for f in fit*.pdb
do
~/rosetta/rosetta_source/bin/rosetta_scripts.linuxgccrelease
\database ~/rosetta/rosetta_database/ -parser:protocol
bbfilt.xml \  extra_res_fa *.params -s "ala.pdb ${f}" -no_optH
false >> bbfilt.log;
done
```

Each conformer passing the filter was visually examined, and those which did not seem realistic, which had internal clashes, or which had conformational flips of the hexose ring were

discarded. For simultaneous conformer sampling, the acceptable conformers were assembled into a single PDB, separated by "TER" records, and the Rosetta topology file (params file) was edited to add the appropriate "PDB_ROTAMERS" line.

Design Runs

The *E. coli* LacI structure was computationally redesigned using Rosetta. The redesign protocols are based off of the enzyme design protocols [8, 26, 27] adapted to the RosettaScripts framework [213].

To maximize diversity in the output sequences, slightly different protocols were employed. The first redesigned the same sets of binding pocket residues for all ligands, in three different permissiveness levels (restrictive, permissive and intermediate, as specified by the resfileslaci_rest.resfile, laci_perm.resfile and laci_int.resfile). Each ligand conformer was designed independently in different runs. Several different design protocols were used for this grouping. The base protocol consisted of an initial Monte Carlo perturbation of the ligand's rigid body position and orientation in an all-alanine binding pocket, followed by selection of a low energy ligand conformer, then design using a scorefunction with a softened LJ repulsive potential and minimization with a standard repulsive scorefunction (2 cycles), followed by design and minimization with a standard repulsive scorefunction (1 cycle), and finally an exhaustive examination of sidechain identities and rotamers in the current context (rah_enzdes.xml). Alternate protocols omitted the initial Monte Carlo pertubation (rah_nojitter.xml), eliminated any ligand rigid body sampling (rah_fixlig.xml), omitted the final exhaustive rotamer examination (rah_nopol.xml), or added more extensive backbone sampling to the regions of the protein in close proximity to the ligand (rah_bbg.xml).

Additionally, an altered version of the scorefunction was used with the base protocol (rah_sum.xml) which replaced the statistical residue pairing energy (fa_pair) with an upweighted Coulombic term, and altered the dG free parameter for the implicit solvation term [214] for arginine nitrogens (-10.0 versus the standard -11.0), sidechain amide nitrogens (-7.8 versus -10.0), and sidechain amide oxygen (-5.85 versus -10.0). These changes had been seen to improve design performance [215] and have since been incorporated into the standard Rosetta scorefunction.

The second set of runs used a more tailored set of design positions, which was further specialized according to ligand (laci.resfile and laci_R.resfile for lactitol and sucralose; laci_6.resfile and laci_6R.resfile for alpha- and beta-fucose). The "_4" and "_6" variants add additional residues which interact with the C4 and C6 positions of the sugar ring, to possibly accommodate the altered C6 position of fucose. The "R" specifies preservation of arginine 197, which interacts with the C2 sugar hydroxyl and may help stabilize the closed conformation.

In contrast to the first set of runs, all accepted conformers of the ligand were considered simultaneously using the standard Rosetta rotamer sampling scheme. The different design protocols were similar to those in the first set (rah2_enzdes.xml, rah2_nojitter.xml, rah2_fixlig.xml, rah2_nopol.xml, rah2_bbg.xml). In addition, these were combinatorial applied to the modified scorefunction (rah2_sum.xml, rah2_sumnojit.xml, rah2_sumfix.xml, rah2_sumnopol.xml, rah2_sumbbg.xml).

Rosetta@Home uses its own job submission system, but the standard Rosetta commandline equivalent to the runs would be:

```
~/rosetta/rosetta_source/bin/minirosetta.linuxgccrelease -
database \ ~/rosetta/rosetta_database -mute all
@2p9h_rah_LigDes_flags \  parser:protocol ${PROTOCOL_XML} -
parser:script_vars \ resfile=${RESFILE};
```

For the modified scorefunction runs, the option "-corrections:chemical:icoor_05_2009 false" was added. All designs produced on Rosetta@Home were kept, without any total score filtering.

Post-processing

The sequences from the Rosetta@Home design runs were extracted from the "silent file"-formatted output files:

```
grep 'ANNOTATED_SEQUENCE:' rah_${RUNTYPE}_silentfile.out | \
rahmakefasta.py > rah_${RUNTYPE}.fas;
```

The structures from Rosetta@Home were rescored:

```
~/rosetta/rosetta_source/bin/rosetta_scripts.linuxgccrelease -
mute \ all @enzscore_flags -parser:protocol enzscore.xml -
extra_res_fa \ ${LIGAND}.params -database
~/rosetta/rosetta_database/ -score:patch \ hack_elec.wts_patch -
out:file:scorefile rah_${RUNTYPE}.sc \  in:file:silent
rah_${RUNTYPE}_silentfile.out;
```

For the modified scoring runs:

```
~/rosetta/rosetta_source/bin/rosetta_scripts.linuxgccrelease -
mute \ all @enzscore_flags -parser:protocol enzscore.xml -
extra_res_fa \ ${LIGAND}.params -database
~/modified_rosetta_database \
corrections:chemical:icoor_05_2009 false -score:patch \
modified.wts_patch -out:file:scorefile rah_${RUNTYPE}.sc \
in:file:silent rah_${RUNTYPE}_silentfile.out;
```

From each run, the resultant structures were filtered on cutoffs based on the median and median absolute deviation. Designs more than 2 MAD worse than the median for any of

total energy, total repulsive energy, total sidechain hydrogen bonding energy, total number of buried unsatisfied hydrogen bonds, ligand interaction energy, ligand repulsive energy, ligand hydrogen bonding energy, or number of ligand hydrogen bonds or with greater than the 90th percentile of buried unsatisfied hydrogen bonds to the ligand were discarded. Cutoffs were made independently for each run type. The 2 MAD cutoff was chosen as it eliminates the extreme worst structures, while keeping those structures which are near "average".

Commands:

```
./gen_enzdes_cutoffs.py -c rescore_cutoffs.mad -o ${RUNTYPE}.cut
\ rah_${RUNTYPE}.sc;
Perl~/rosetta/rosetta_source/src/apps/public/enzdes/DesignSelect
.pl -d rah_${RUNTYPE}.sc -c rah_${RUNTYPE}.cut -tag_column last
> \ rah_${RUNTYPE}_filt.sc;
./filter_fas.py rah_${RUNTYPE}.fas `awk '{ print $NF }' \
rah_${RUNTYPE}_filt.sc` > rah_${RUNTYPE}_filt.fas;
```

The unique filtered designs from all the Rosetta@Home runs were assembled on a per-ligand basis. (For fucose, both alpha and beta fucose runs were combined.)

Commands:

```
./uniquify_fas.py -b rah_*_filt.fas>all_sequences.fas;
```

As the experimental synthesis scheme is limited in the length of the DNA segments which it can produce, the designed sequences were segmented into three or four separate regions, which encompass most of the variability in the designs, but are each small enough to be synthesized. Sequence modifications were imposed to remove mutations deemed to be spurious by visual inspection. Two sets of reversions were used, a more extensive ligand-specific reversion and a less extensive, ligand-independent one.

Commands:

```
./segment_fas.py -m strong_${LIGAND}.revert --tag _strong --
cutoff \ 10 -o 62 all_sequences.fas 73-125 148-197 245-296;
./segment_fas.py -m mild.revert   --tag _mild   --cutoff  5 -o
62 \ all_sequences.fas 73-125 148-197 245-296;
```

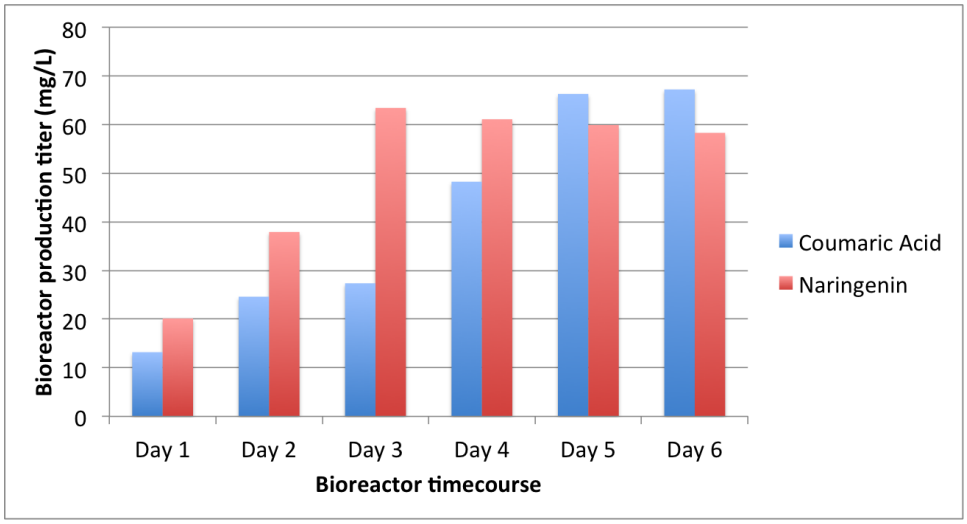
The separately segmented files were then combined and the unique sequences we selected, preferring the sequences from the more extensive ("strong") reversion scheme, and those which have a greater representation in the full set.

Commands:

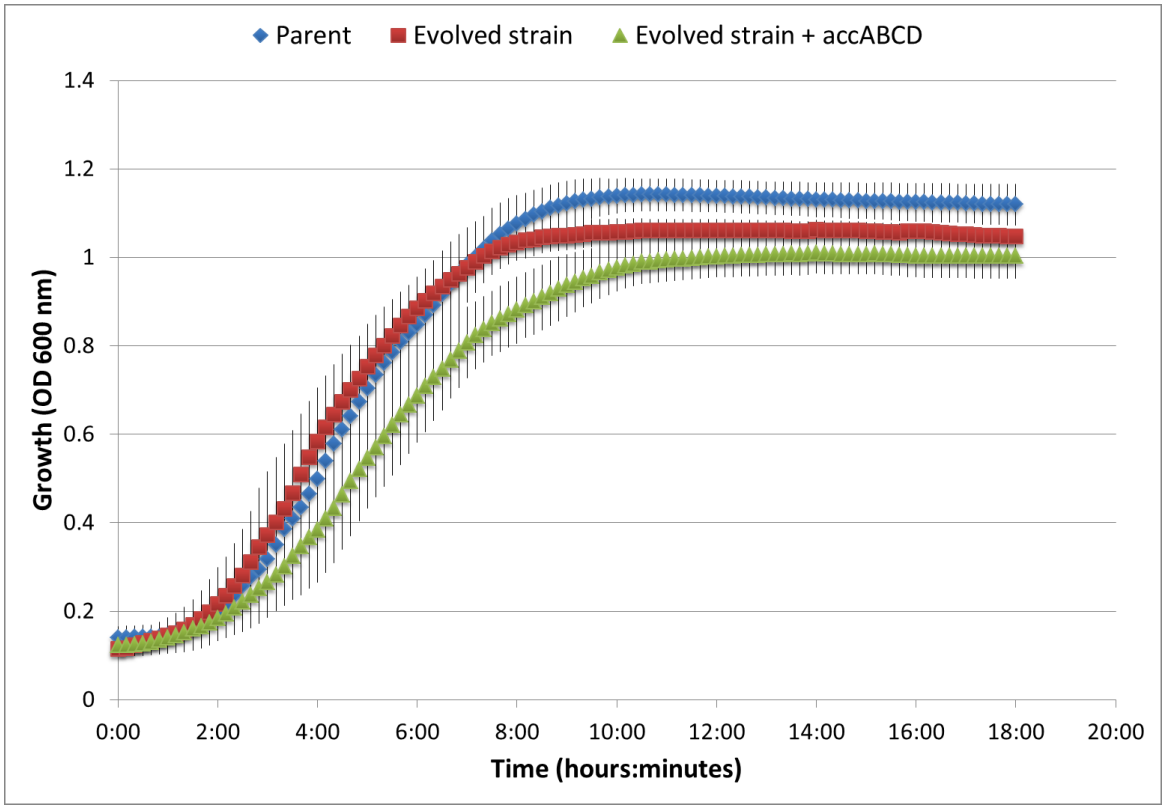
```
./uniquify_fas.py all_sequences.fas_strong_73-125.fas  \
at_home_all.fas_mild_73-125.fas  --keeporder | head -n 600  > \
sucralose_73-125.fas;
./uniquify_fas.py all_sequences.fas_strong_148-197.fas \
at_home_all.fas_mild_148-197.fas --keeporder | head -n 3650 > \
sucralose_148-197.fas;
./uniquify_fas.py all_sequences.fas_strong_245-296.fas \
at_home_all.fas_mild_245-296.fas --keeporder | head -n 3650 > \
sucralose_245-296.fas;
```

Supplementary Figures

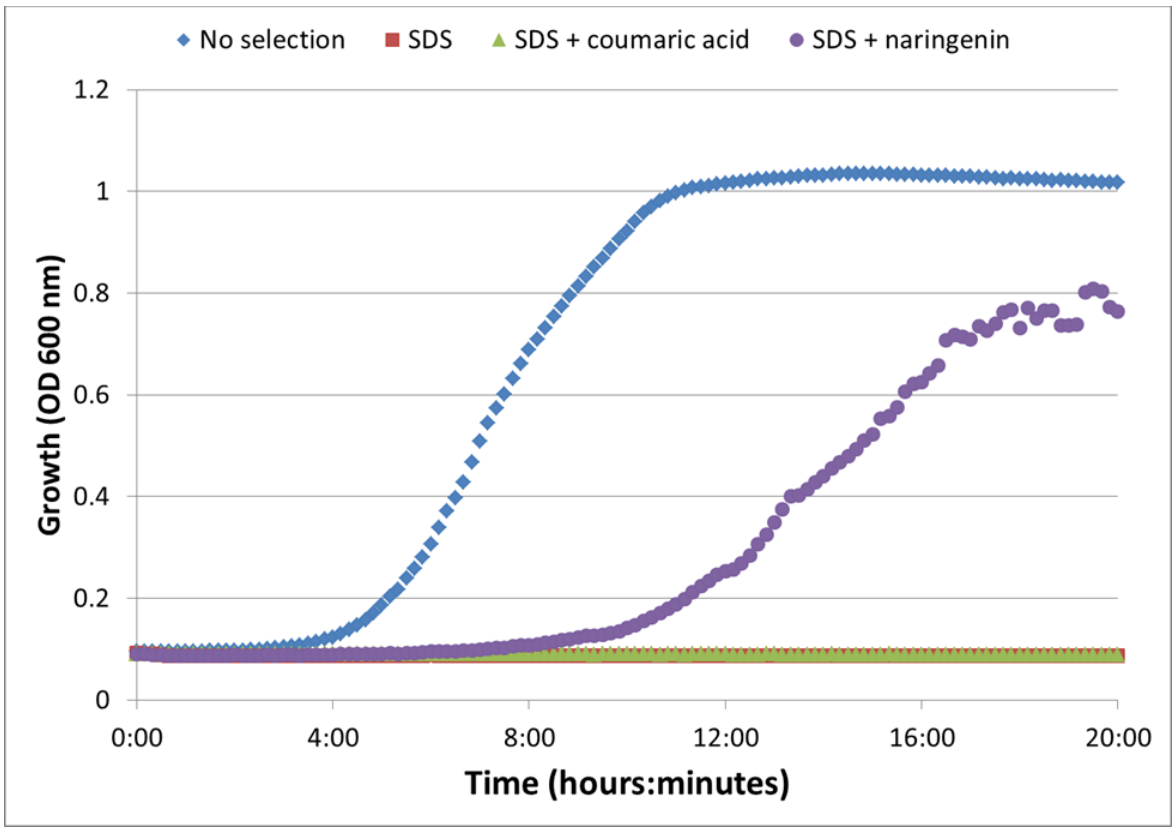
Supplementary Figures for Chapter 2



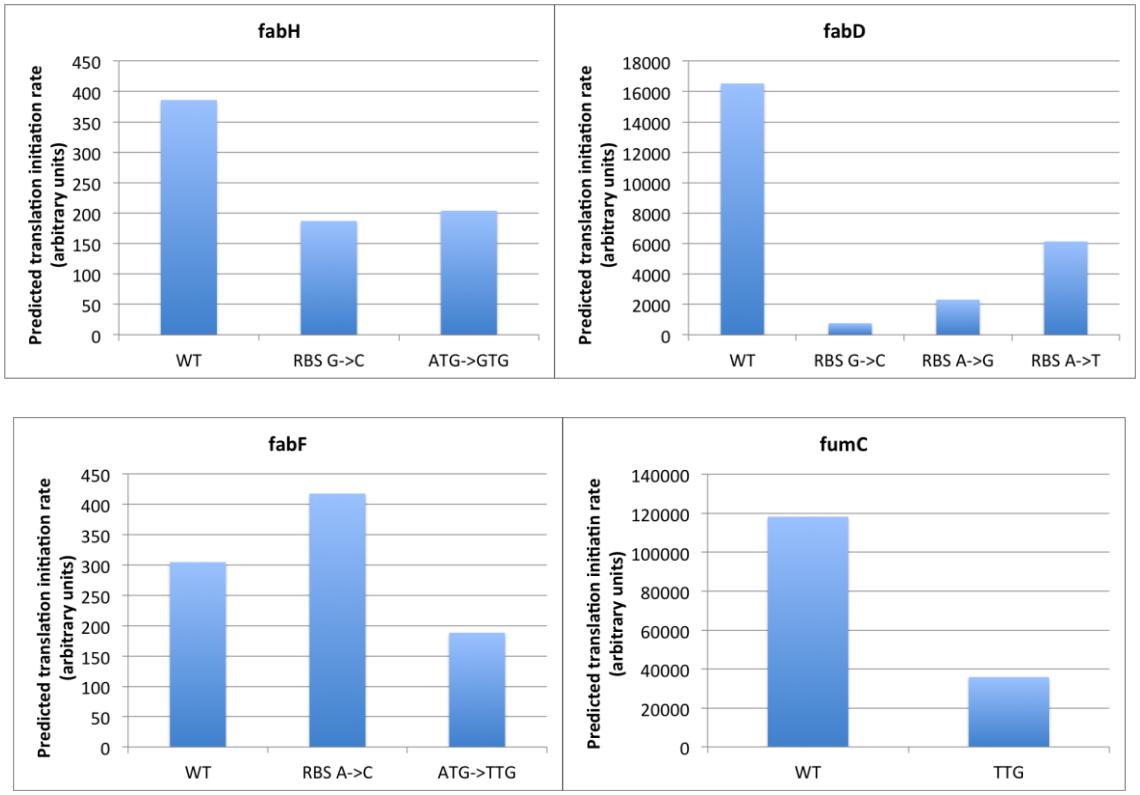
Supplementary Figure S2.1
Time course of naringenin and coumaric acid production titers in bioreactor.



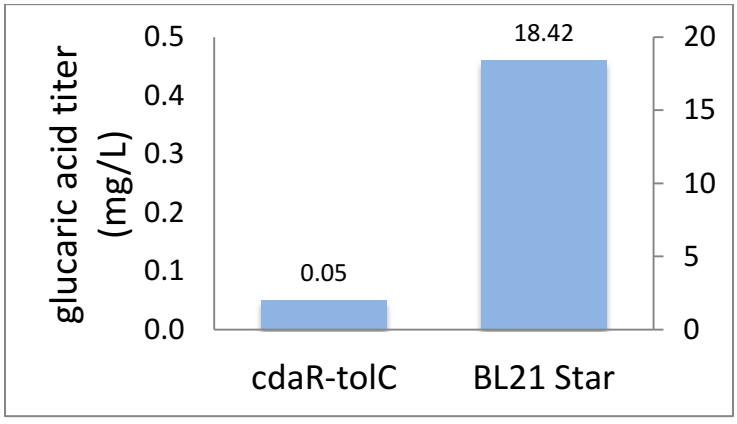
Supplementary Figure S2.2
Growth comparison of TtgR-TolC parent strain, evolved strain, and evolved strain expressing accABCD genes from a plasmid. Mean and standard deviation of three replicates.



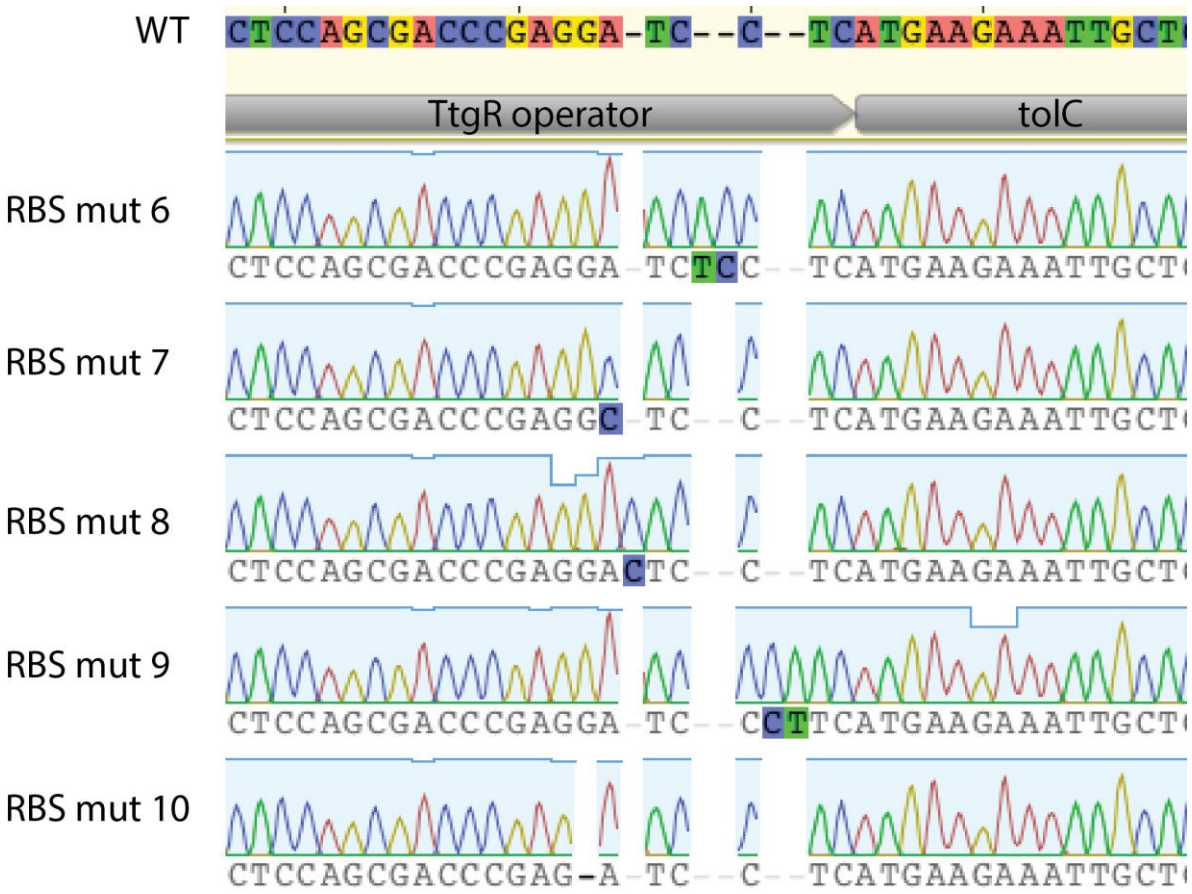
Supplementary Figure S2.3
TtgR-TolC strain growth comparison under SDS selection, exposed to 50 µM naringenin or 50 µM coumaric acid.



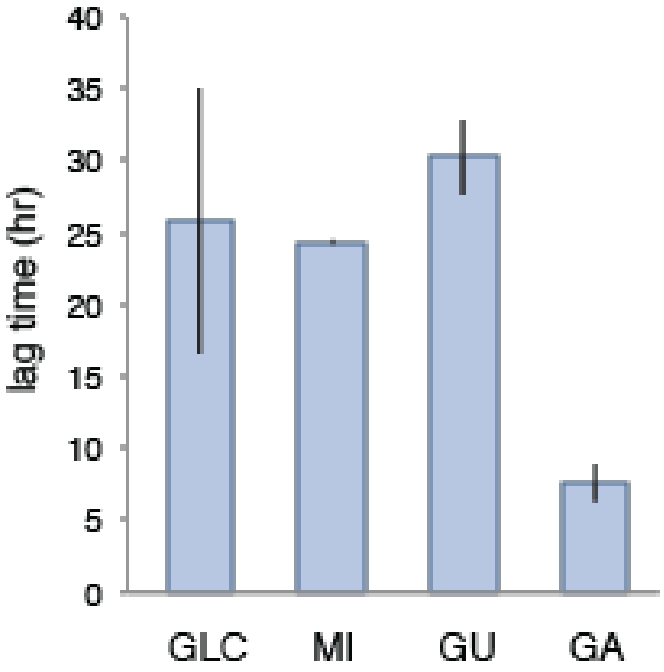
Supplementary Figure S2.4
Predicted translation initiation rate of RBS mutants.



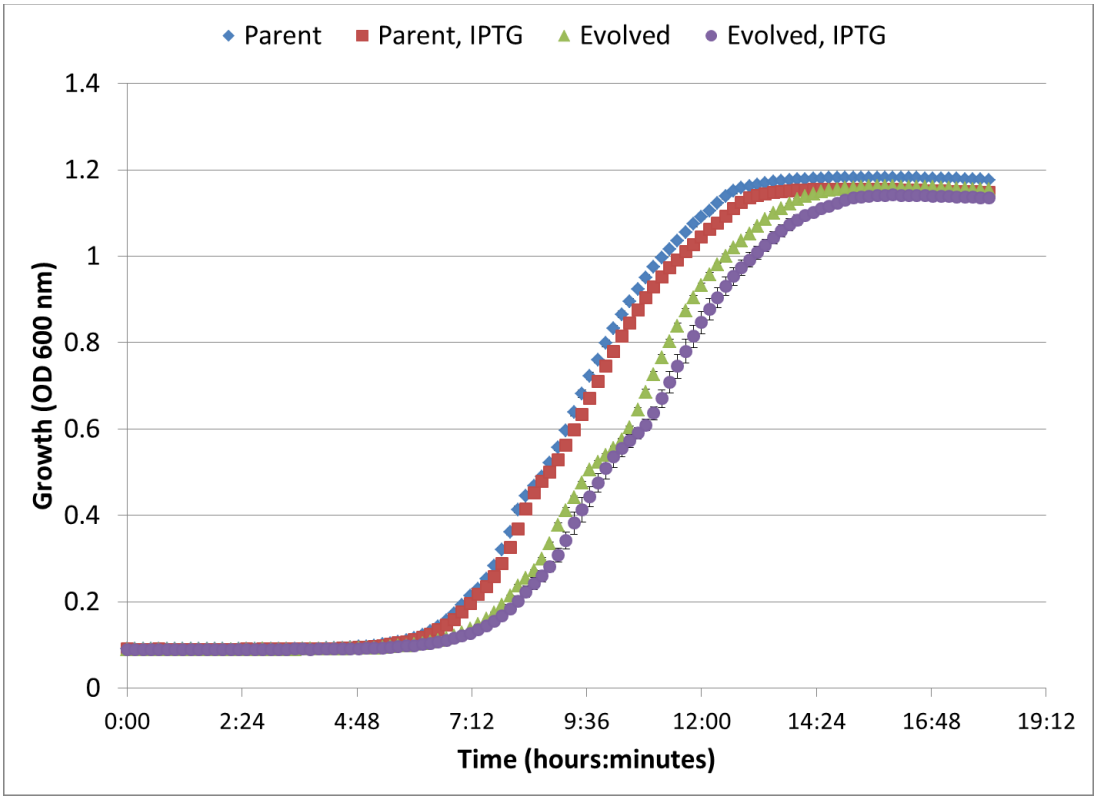
Supplementary Figure S2.5
Glucaric acid production by the pT7GAEXP plasmid in the CdaR-TolC sensor-selector strain (*E. coli* K12 derivative) and BL21 Star.



Supplementary Figure S2.6
TtgR operator RBS spacing modifications.



Supplementary Figure S2.7
CdaR-tolC sensor selector strain response to glucaric acid pathway intermediates. Abbreviations: GLC, glucose; MI, myo-inositol; GU, glucuronic acid; GA, glucaric acid.

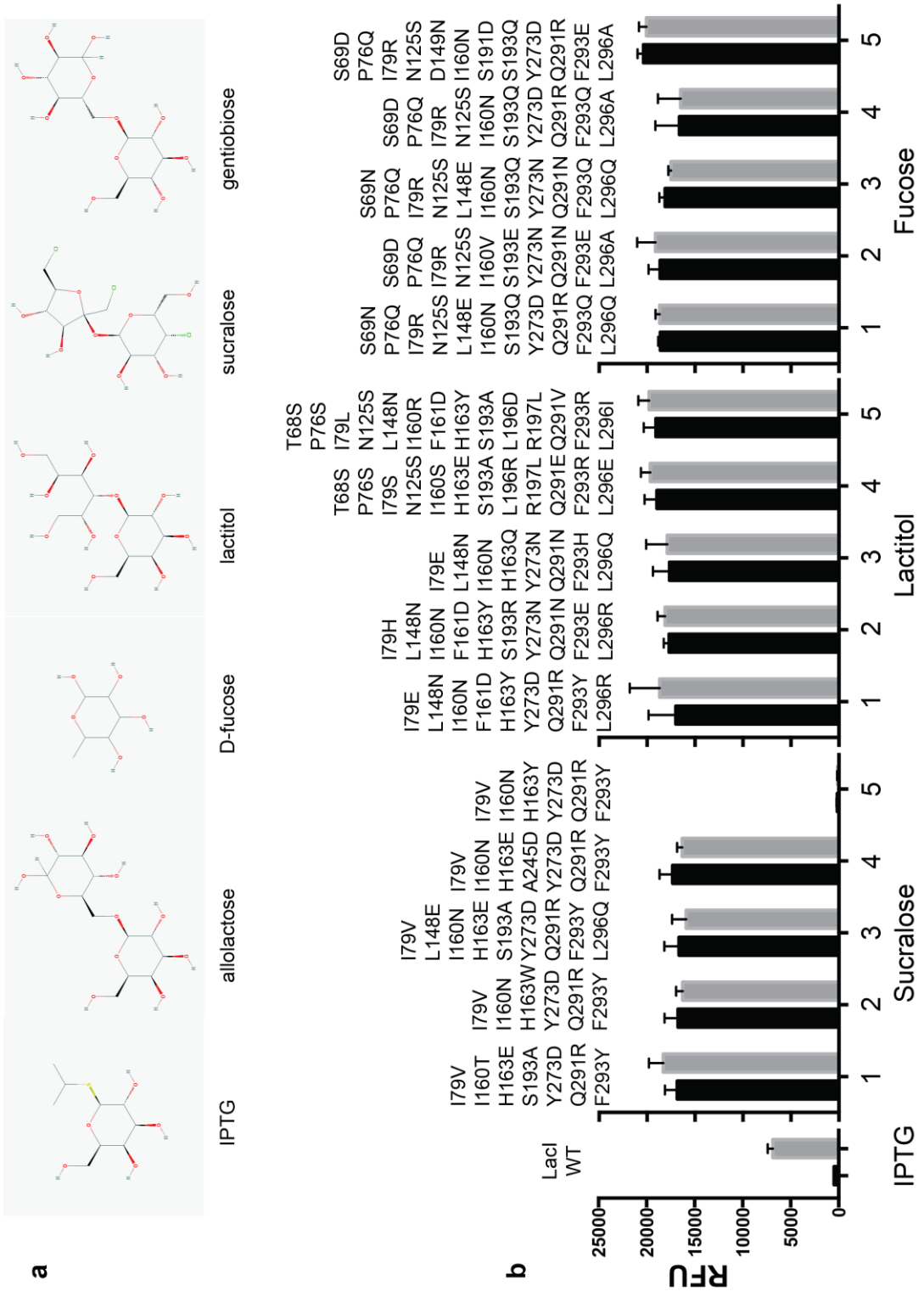


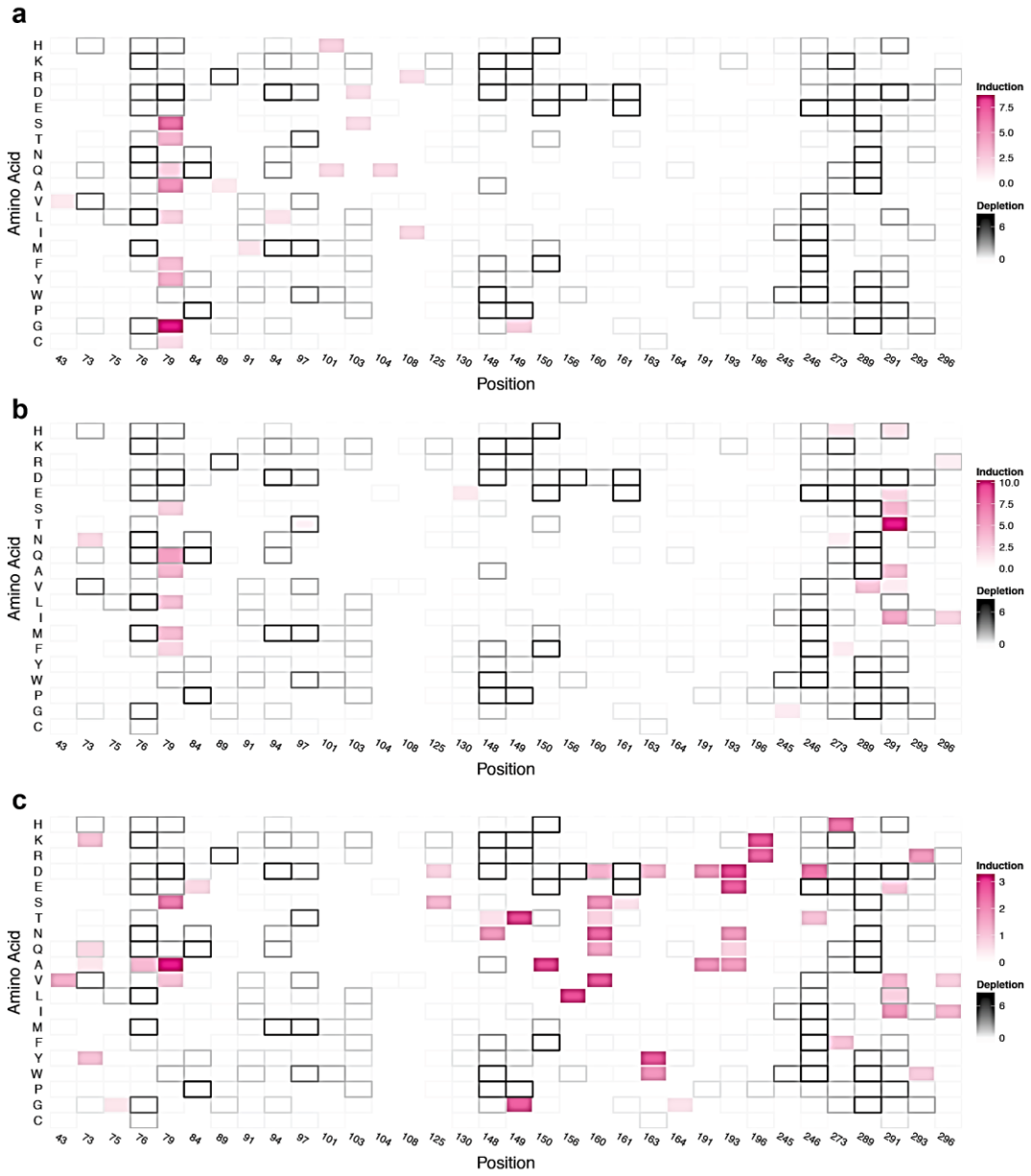
Supplementary Figure S2.8
Growth comparison of CdaR-TolC parent and best evolved glucaric acid production strain. IPTG used to induce pT7GAEXP glucaric acid production plasmid. Mean and standard deviation of five replicates.

Supplementary Figures for Chapter 3

Supplementary Figure S3.1 Chemical structure of ligands and induction of top-five scoring full-length Rosetta design variants. **(a)** Chemical structure of allolactose and IPTG (native and synthetic inducers of LacI, respectively), and the four target inducers: fucose, lactitol, sucralose and gentiobiose. **(b)** Induction response in relative fluorescence units (RFU) with and without ligand for WT-LacI and top-five scoring full-length Rosetta design variants for sucralose, lactitol and fucose. WT-LacI was induced with IPTG, and the full-length Rosetta design variants were induced with their respective target ligands. The mutations in each Rosetta-designed variant is shown above the bar graph. All ligands were added at 10 mM concentration.

Supplementary Figure S3.1 (Continued)

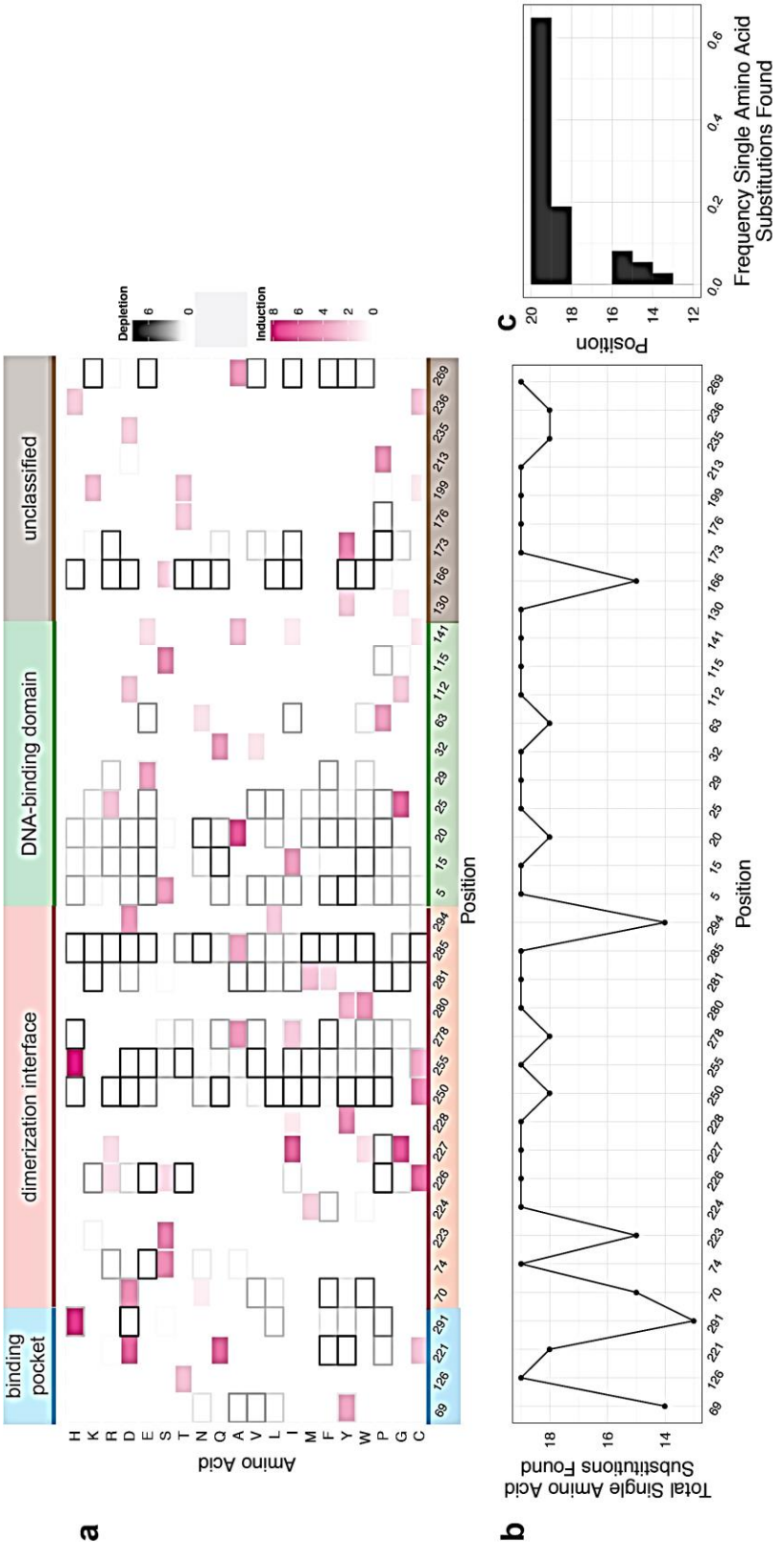


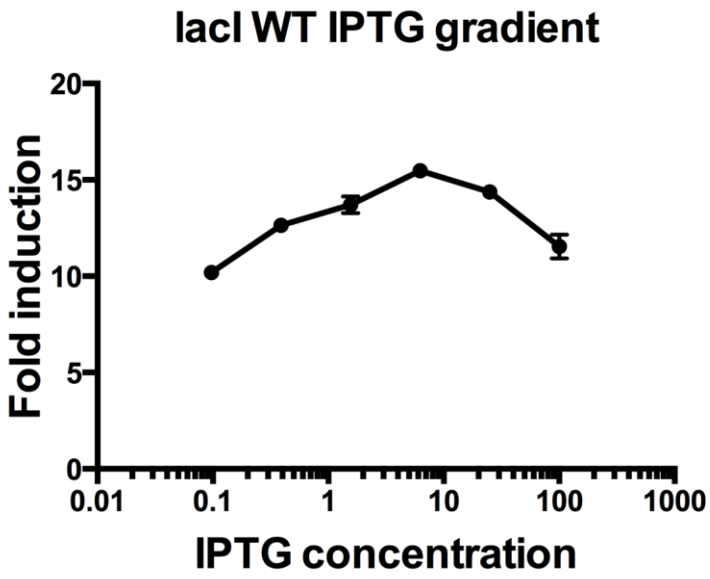


Supplementary Figure S3.2 Comparison of fucose, lactitol, and sucralose response versus single-amino acid substitutions found after negative selection. **(a)** Fucose responsive induction values are shown pink. The induction values show the maximum weigthed fold-change of response after positive selection. The black outlines indicate depletion of next-generation sequencing reads for single amino-acid substitutions after negative selection. The depletion value is the $\log_2$ fold-change of reads prior to negative selection divided by the read counts after negative selection. Higher depletion values indicate position and side-chain combinations that are lost after negative selection. Read counts were quantile normalized between pre- and post-selection separately for each amplicon (see Supplemental Methods). **(b)** Lactitol responsive induction values versus depletion values. **(c)** Sucralose responsive induction values versus depletion values. Negative depletion values are not shown.

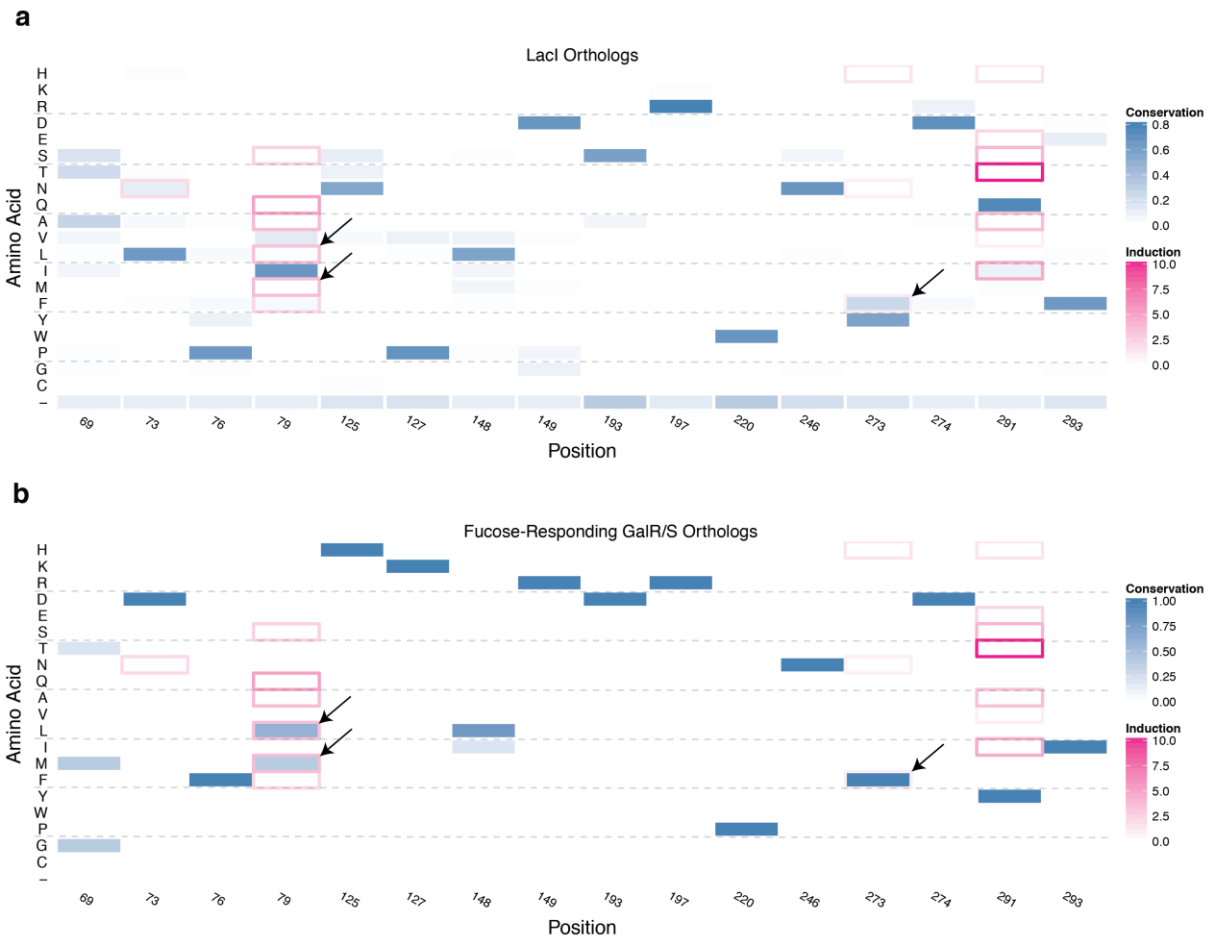
Supplementary Figure S3.3 Comparison of gentiobiose response versus single-amino acid substitutions found after negative selection. **(a)** Induction values for gentiobiose-responding mutants are shown in pink. The induction values show the maximum weighted fold-change of response after positive selection. The color shades outside the amino acid substitution profile denotes the location of the residue in the binding pocket, dimerization interface, DNA-binding domain or as unclassified. The black outlines indicate depletion of next-generation sequencing reads for single amino-acid substitutions. The depletion value is the log2 fold-change of reads prior to negative selection divided by the read counts after negative selection. Read counts were quantile normalized between pre- and post-selection separately for each amplicon (see Supplemental Methods). **(b)** For each and every position of wild-type LacI, 19 substitutions were synthesized for the single amino-acid substitution library. By next-generation sequencing we measured how many of the 19 possible substitutions were found either before or after negative selection for positions showing response to gentiobiose. **(c)** Most (>80%) of the positions harboring gentiobiose response were found to have at least 18 of 19 single amino-acid substitutions prior to positive selection. For all 360 positions of LacI (not shown), we found 195 (~54%) positions contained all 19 substitutions, 238 (~66%) contained at least 18, and 306 (85%) contained at least 14 substitutions.

Supplementary Figure S3.3 (Continued)





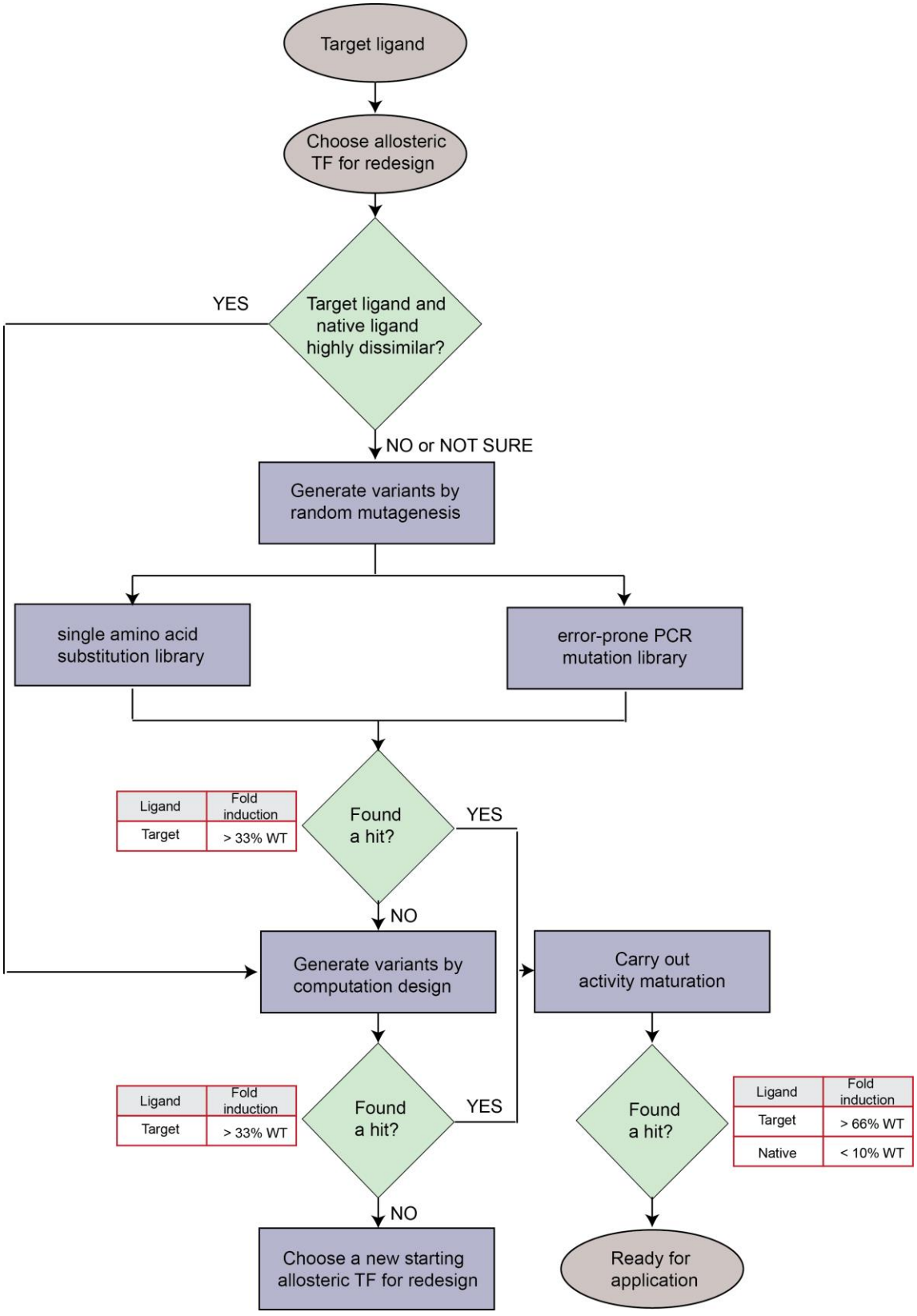
Supplementary Figure S3.4 Fold induction of WT LacI with IPTG. Dose-response curve of WT LacI with IPTG, fold induction measured by FACS shown on Y-axis and IPTG concentration on the X-axis.

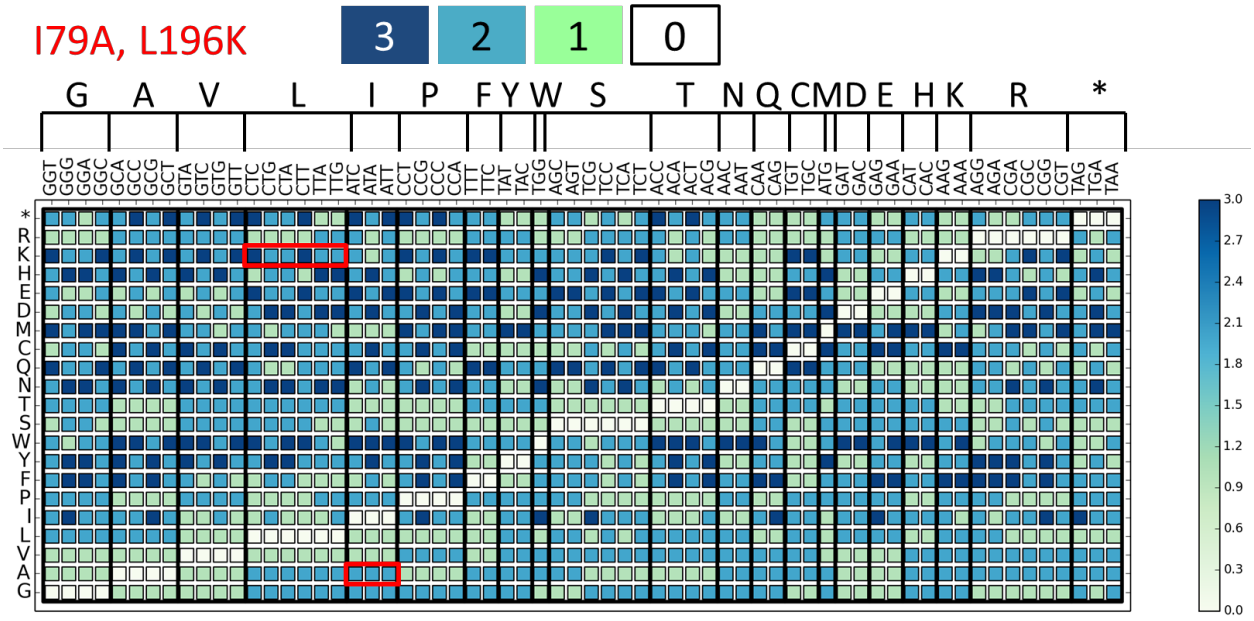


Supplementary Figure S3.5 Comparison of conservation of amino acids and mutations found for fucose response. **(a)** A set of 41 LacI orthologs were aligned and the frequency of amino acid utilization is shown in blue. Mutations conferring fucose repsonse are shown as red outlines. **(b)** Five experimentally validated sequences of GalR/S known to bind fucose were aligned with E. coli LacI and shown with respect to LacI positions. Mutations at positions 79 and 273 overlap with preferentially conserved amino acids in the GalR set, shown with arrows. The highest inducer at position 291 was conserved in neither LacI nor fucose-responding GalR/S.

Supplementary Figure S3.6 User guide to allosteric TF redesign**.** A detailed flowchart that guides the user through the choice of mutagenesis methods based on the choice of the target ligand for allosteric TF redesign. We offer general guidelines on what we consider acceptable fold induction and specificity values by target and native ligands, presented as proportion of WT aTF induction, after the two-stage enrichment screen and following activity maturation. These guidelines could be adjusted on a case-by-case basis depending on the number and quality of ligand-responsive variants after the two-stage enrichment screen, and the nature of downstream application.

Supplementary Figure S3.6 (Continued)





Supplementary Figure S3.7 Minimum Hamming distances from each codon to any codon encoding each amino acid. On the horizontal axis, each codon is displayed, grouped by the residue it encodes. On the vertical axis, each residue is displayed. Each small box represents the Hamming distance between that codon and any codon that encodes that amino acid; the boxes are color coded by this distance: 3, dark blue; 2, light blue, 1, green, 0, white. For example, a Hamming distance of two means that two separate single nucleotide changes would be required to mutate to any codon encoding the given amino acid. In red boxes, the mutations required to mutate LacI codon Ile79 to Alanine, or Leu196 to Lysine, are highlighted.

Supplementary Tables

Supplementary Tables for Chapter 2

Supplementary Table S2.1
Sensor-promoter pairs used in this study.

Sensor	**Cognate promoter**
BenM	benAp [56]
theoRR	None: riboswitch [70]
btuB	None: riboswitch [48]
AlkS	alkBp [59]
LacI	pLlacO [45]
XylR	xylAp [97]
CdaR	gudPp [53]
MphR	mphAp [50]
TetR	pLtetO [45]
TtgR	ttgAp [58]

Supplementary Table S2.2
List of codon optimized gene sequences used in this study.

Name	Sequence 5' to 3'
ttgR	ATGGTGCGTCGCACCAAAGAAGAAGCACAGGAAACGCGTGCGCAGATTATCGAAGCGGCCGAACGCGCG TTTTATAAACGTGGTGTGGCACGTACCACGCTGGCAGATATTGCAGAACTGGCAGGTGTTACCCGCGGT GCAATCTACTGGCATTTCAACAATAAAGCCGAACTGGTTCAGGCACTGCTGGATTCTCTGCACGAAACG CATGATCACCTGGCCCGTGCAAGCGAATCTGAAGATGAACTGGACCCGCTGGGCTGCATGCGCAAACTG CTGCTGCAGGTGTTTAACGAACTGGTTCTGGATGCACGTACCCGTCGCATTAATGAAATCCTGCATCAC AAATGCGAATTTACGGATGATATGTGTGAAATTCGTCAGCAGCGCCAGAGCGCCGTGCTGGATTGTCAT AAAGGTATCACCCTGGCACTGGCAAACGCAGTTCGTCGCGGTCAGCTGCCGGGTGAACTGGATGTGGAA CGCGCAGCGGTTGCGATGTTTGCCTATGTGGATGGCCTGATTGGTCGTTGGCTGCTGCTGCCGGATAGT GTTGATCTGCTGGGCGATGTGGAAAAATGGGTTGATACCGGTCTGGATATGCTGCGTCTGAGCCCGGCG CTGCGCAAATAA
mphR	ATGCCGCGTCCGAAACTGAAATCTGACGACGAAGTTCTGGAAGCGGCGACCGTTGTTCTGAAACGTTGC GGTCCGATCGAATTCACCCTGTCTGGTGTTGCGAAAGAAGTTGGTCTGTCTCGTGCGGCGCTGATCCAG CGTTTCACCAACCGTGACACCCTGCTGGTTCGTATGATGGAACGTGGTGTTGAACAGGTTCGTCACTAC CTGAACGCGATCCCGATCGGTGCGGGTCCGCAGGGTCTGTGGGAATTCCTGCAGGTTCTGGTTCGTTCT ATGAACACCCGTAACGACTTCTCTGTTAACTACCTGATCTCTTGGTACGAACTGCAGGTTCCGGAACTG CGTACCCTGGCGATCCAGCGTAACCGTGCGGTTGTTGAAGGTATCCGTAAACGTCTGCCGCCGGGTGCG CCGGCGGCGGCGGAACTGCTGCTGCACTCTGTTATCGCGGGTGCGACCATGCAGTGGGCGGTTGACCCG GACGGTGAACTGGCGGACCACGTTCTGGCGCAGATCGCGGCGATCCTGTGCCTGATGTTCCCGGAACAC GACGACTTCCAGCTGCTGCAGGCGCACGCGTAA
MIOX	GTGAAAGTGGATGTTGGCCCGGACCCGAGCCTGGTTTACCGCCCGGATGTGGACCCGGAAATGGCAAAA AGCAAAGATTCGTTTCGTAACTACACCAGTGGCCCGCTGCTGGATCGTGTTTTTACCACGTATAAACTG ATGCATACCCACCAGACGGTTGACTTTGTCAGCCGTAAACGCATTCAATATGGCGGTTTCTCTTACAAG AAAATGACCATCATGGAAGCGGTGGGCATGCTGGATGACCTGGTTGATGAATCAGATCCGGACGTCGAT TTTCCGAATTCGTTTCATGCGTTCCAGACGGCCGAAGGTATTCGCAAAGCCCACCCGGACAAAGATTGG TTCCATCTGGTCGGCCTGCTGCACGATCTGGGTAAAATCATGGCACTGTGGGGTGAACCGCAGTGGGCT GTGGTTGGTGATACCTTTCCGGTGGGTTGCCGTCCGCAAGCAAGTGTCGTGTTTTGTGACTCCACCTTC CAGGACAACCCGGATCTGCAAGACCCGCGCTATTCAACGGAACTGGGCATGTACCAGCCGCATTGCGGT CTGGAAAACGTGCTGATGTCGTGGGGTCACGATGAATACCTGTACCAGATGATGAAATTCAACAAATTC AGCCTGCCGTCTGAAGCCTTCTACATGATCCGTTTCCATAGTTTCTACCCGTGGCACACCGGCGGTGAT TATCGCCAGCTGTGCTCCCAGCAAGACCTGGATATGCTGCCGTGGGTGCAAGAATTCAACAAATTCGAT CTGTACACGAAATGTCCGGATCTGCCGGACGTTGAATCTCTGCGTCCGTACTACCAAGGTCTGATTGAT AAATACTGTCCGGGCACCCTGTCGTGGTAA

Supplementary Table S2.3
Genomic MAGE targets in each cycle naringenin pathway diversification.

Evolution round	Up-regulation targets	Down-regulation targets	Coding targets
1	*accABCD aceEF gapA lpdA pgk*	*acnA fumC mdh sucC*	
2	*accABCD aceEF gapA lpdA pgk*	*acnA fumC mdh sucC*	
3	*accABCD aceEF gapA lpdA pgk*	*acnA fumC mdh sucC*	
4		*fabBDFH*	*tyrA*_M53I *tyrA*_A354V *aroG*_D146N

Supplementary Table S2.4

Oligonucleotides used for naringenin mutagenesis. * symbol indicates phosphorthioate bonds.

Name	Function	Sequence
accA_T7	promoter insertion	C*A*A*G*GAAATTCAGACTCATAGTATTCCTGTATTATCTCCCTATAG TGAGTCGTATTAGTCAAACTCCAGTTCCACCTGCTCCGAACCAAT
accA_JC23100	promoter insertion	A*T*T*C*AGACTCATAGTATTCCTGTATTAGCTAGCACTGTACCTAGG ACTGAGCTAGCCGTCAAGTCAAACTCCAGTTCCACCTGCTCCGAA
accB_T7	promoter insertion	C*G*G*T*GAAACGCCTGTCACAATCACACTAAACAATAATACGACTCA CTATAGGGAGAAGAGTACGGAACCCACTCATGGATATTCGTAAGA
accB_JC23100	promoter insertion	A*A*C*G*CCTGTCACAATCACACTAAACAATTGACGGCTAGCTCAGTC CTAGGTACAGTGCTAGCAGAGTACGGAACCCACTCATGGATATTC
accD_T7	promoter insertion	G*T*T*C*AATCCAGCTCATTAGGGACCTTTCTGTCTTCTCCCTATAGT GAGTCGTATTAGAACCTGGTTCGATGCCAGTTTTATCTTTGGGGA
accD_JC23100	promoter insertion	T*C*C*A*GCTCATTAGGGACCTTTCTGTCTGCTAGCACTGTACCTAGG ACTGAGCTAGCCGTCAAGAACCTGGTTCGATGCCAGTTTTATCTT
aceE_T7	promoter insertion	A*A*C*G*TTCTGACATGGGTTATTCCTTATCTATCTTCTCCCTATAGT GAGTCGTATTAAATAACGTTGAGTTTTCTGGAACCTGTCCCATTG
aceE_JC23100	promoter insertion	C*T*G*A*CATGGGTTATTCCTTATCTATCTGCTAGCACTGTACCTAGG ACTGAGCTAGCCGTCAAAATAACGTTGAGTTTTCTGGAACCTGTC
lpdA_T7	promoter insertion	T*T*T*G*ATTTCAGTACTCATCATGACCTCTATATATCTCCCTATAGT GAGTCGTATTATTTATCTCCGGCGGTCATACCCGTCGTCTTTCAG
lpdA_JC23100	promoter insertion	T*T*C*A*GTACTCATCATGACCTCTATATAGCTAGCACTGTACCTAGG ACTGAGCTAGCCGTCAATTTATCTCCGGCGGTCATACCCGTCGTC
gapA_T7	promoter insertion	G*T*A*A*TTTTACAGGCAACCTTTTATTCACTAACATAATACGACTCA CTATAGGGAGAAATAGCTGGTGGAATATATGACTATCAAAGTAGG
gapA_JC23100	promoter insertion	T*T*A*C*AGGCAACCTTTTATTCACTAACATTGACGGCTAGCTCAGTC CTAGGTACAGTGCTAGCAATAGCTGGTGGAATATATGACTATCAA
pgk_T7	promoter insertion	A*T*C*T*TAATTACAGACATGGTGAATCCTCTCGTTTCTCCCTATAGT GAGTCGTATTAGATTCTAAAAGTTTTGCAGACGCTGCTTGCGTCT
pgk_JC23100	promoter insertion	A*T*T*A*CAGACATGGTGAATCCTCTCGTTGCTAGCACTGTACCTAGG ACTGAGCTAGCCGTCAAGATTCTAAAAGTTTTGCAGACGCTGCTT
fumB_BTG	start codon degeneracy	A*C*G*C*CATTTTCGAATAACAAATACAGAGTTACAGGCTGGAAGCTB TGTCAAACAAACCCTTTATCTACCAGGCACCTTTCCCGATGGGGA
fumC_BTG	start codon degeneracy	C*A*T*C*AATCGCCCCCATCGAATCTTTTTCGCTGCGTACTGTATTCA VGACCTGCTCCTCACCTGATTAATTTTTTCTTTCTGTTTTGCTTT
mdh_BTG	start codon degeneracy	C*G*C*C*TGGCCAATACCGCCAGCAGCGCCGAGGACTGCGACTTTCAV CCTAAACTCCTTATTATATTGATAAACTAAGATATGTTGCTCCGC
acnA_BTG	start codon degeneracy	C*T*G*A*AGAGAATCAGGGCTTCGCAACCCTGTCATTAAGGAGGAGCT BTGTCGTCAACCCTACGAGAAGCCAGTAAGGACACGTTGCAGGCC
tyrA_Met53Ile	coding mutation	G*T*A*C*ACCCAGAGCTTCCGCCTCTGCACGACGCGAGGCCAAAATAG ATGCCTCGCGCTCCGGAACATAAATAGGCAGTCCAAAGCGGCTTT
tyrA_Ala354Val	coding mutation	T*C*A*T*TCGCCTGACGCAATAACACGCGGCTTTCACTCTGAAAACGC TGTACGTAATCGCCGAACCAGTGCTCCACCTTGCGGAAACTGTCA
aroG_Asp146Asn	coding mutation	C*C*A*G*CTCATCAGGTCAGCGAGATATTGTGGGGTGATCATATTGAG AAACTCACCTGCCGCTGGCAGACCGCTGTCGTTAATATCAAGCAG
accA_RBS	RBS degeneracy	C*A*A*T*CGGCTGTTCAAAATCAAGGAAATTCAGACTCATAGTHHHHY YYYYHHTTAGTCAAACTCCAGTTCCACCTGCTCCGAACCAATGAG
accB_RBS	RBS degeneracy	C*G*G*T*GAAACGCCTGTCACAATCACACTAAACAAAGAGDDRRRRRD DDDCTCATGGATATTCGTAAGATTAAAAAACTGATCGAGCTGGTT
accC_RBS	RBS degeneracy	C*C*G*G*TAGAATTTGACGAGCCGCTGGTCGTCATCGAGTAADDRRRR RDDDDAACATGCTGGATAAAATTGTTATTGCCAACCGCGGCGAGA
accD_RBS	RBS degeneracy	G*G*T*G*GGAGTAATGTTGCTTTTAATTCGTTCAATCCAGCTCATTAG HHHHYYYYYHHTCTGAACCTGGTTCGATGCCAGTTTTATCTTTGG
aceE_RBS	RBS degeneracy	C*G*A*T*CGGATCCACGTCATTTGGGAAACGTTCTGACATGGGHHHHY YYYYHHCTATCTAATAACGTTGAGTTTTCTGGAACCTGTCCCATT

Supplementary Table S2.4 (Continued)

aceF_RBS	RBS degeneracy	C*C*G*A*TGTCCGGTACTTTGATTTCGATAGCCATTATHHHHYYYYYH HTTACGCCAGACGCGGGTTAACTTTATCTGCATCGATGTTGAATT
lpd_RBS	RBS degeneracy	C*C*C*A*AGTACCACGACCTGAGTTTTGATTTCAGTACTCATCATHHH HYYYYYHHATTTATCTCCGGCGGTCATACCCGTCGTCTTTCAGGC
gapA_RBS	RBS degeneracy	T*T*T*G*TAATTTTACAGGCAACCTTTTATTCACTAACAAATDDRRRR RDDDDTATATGACTATCAAAGTAGGTATCAACGGTTTTGGCCGTA
pgk_RBS	RBS degeneracy	A*G*C*A*AGATCCAGATCGGTCATCTTAATTACAGACATGGTHHHHYY YYYHHTTGATTCTAAAAGTTTTGCAGACGCTGCTTGCGTCTTACC
sucC_BTG	start codon degeneracy	A*T*A*G*CGGGCAAAAAGTTGTTTTGCCTGATATTCATGTAAGTTCAV GTGTTCTGTCCATCCTTCAGTAATCGTTATCTTTTAAACCGTAGA
fabB_mut1	RBS degeneracy	A*A*C*A*ATGCCCAGGCCAGTAATCACTGCACGTTTCATTCAATACCT CNGTAAGTCGCACATAGAGTAAGTTTCGAATGCACAATAGCGTAC
fabB_mut2	RBS degeneracy	A*A*C*A*ATGCCCAGGCCAGTAATCACTGCACGTTTCATTCAATACCT NTGTAAGTCGCACATAGAGTAAGTTTCGAATGCACAATAGCGTAC
fabB_mut3	RBS degeneracy	A*A*C*A*ATGCCCAGGCCAGTAATCACTGCACGTTTCATTCAATACCN CTGTAAGTCGCACATAGAGTAAGTTTCGAATGCACAATAGCGTAC
fabB_mut4	RBS degeneracy	A*A*C*A*ATGCCCAGGCCAGTAATCACTGCACGTTTCATTCAATACNT CTGTAAGTCGCACATAGAGTAAGTTTCGAATGCACAATAGCGTAC
fabB_mut5	RBS degeneracy	A*A*C*A*ATGCCCAGGCCAGTAATCACTGCACGTTTCATTCAATANCT CTGTAAGTCGCACATAGAGTAAGTTTCGAATGCACAATAGCGTAC
fabF_mut1	RBS degeneracy	T*G*C*C*CAGTCCGGTCACAACTACACGACGCTTAGACACGTTTGTCC TCNAGGGAGGGAAAAAATGATTCTAGTGGGACAAAAAGATAAAAC
fabF_mut2	RBS degeneracy	T*G*C*C*CAGTCCGGTCACAACTACACGACGCTTAGACACGTTTGTCC TNCAGGGAGGGAAAAAATGATTCTAGTGGGACAAAAAGATAAAAC
fabF_mut3	RBS degeneracy	T*G*C*C*CAGTCCGGTCACAACTACACGACGCTTAGACACGTTTGTCC NCCAGGGAGGGAAAAAATGATTCTAGTGGGACAAAAAGATAAAAC
fabF_mut4	RBS degeneracy	T*G*C*C*CAGTCCGGTCACAACTACACGACGCTTAGACACGTTTGTCN TCCAGGGAGGGAAAAAATGATTCTAGTGGGACAAAAAGATAAAAC
fabF_mut5	RBS degeneracy	T*G*C*C*CAGTCCGGTCACAACTACACGACGCTTAGACACGTTTGTNC TCCAGGGAGGGAAAAAATGATTCTAGTGGGACAAAAAGATAAAAC
fabF_mut6	RBS degeneracy	T*G*C*C*CAGTCCGGTCACAACTACACGACGCTTAGACACGTTTGNCC TCCAGGGAGGGAAAAAATGATTCTAGTGGGACAAAAAGATAAAAC
fabD_mut1	RBS degeneracy	C*C*C*T*GTCCAGGGAACACAAATGCAAATTGCGTCATGTTTTAATCC TNATCCTAGAAACGAACCAGCGCGGAGCCCCAGGTGAATCCACCG
fabD_mut2	RBS degeneracy	C*C*C*T*GTCCAGGGAACACAAATGCAAATTGCGTCATGTTTTAATCC NTATCCTAGAAACGAACCAGCGCGGAGCCCCAGGTGAATCCACCG
fabD_mut3	RBS degeneracy	C*C*C*T*GTCCAGGGAACACAAATGCAAATTGCGTCATGTTTTAATCN TTATCCTAGAAACGAACCAGCGCGGAGCCCCAGGTGAATCCACCG
fabD_mut4	RBS degeneracy	C*C*C*T*GTCCAGGGAACACAAATGCAAATTGCGTCATGTTTTAATNC TTATCCTAGAAACGAACCAGCGCGGAGCCCCAGGTGAATCCACCG
fabD_mut5	RBS degeneracy	C*C*C*T*GTCCAGGGAACACAAATGCAAATTGCGTCATGTTTTAANCC TTATCCTAGAAACGAACCAGCGCGGAGCCCCAGGTGAATCCACCG
fabH_mut1	RBS degeneracy	C*A*G*A*TAGCTGCCAGTACCAATAATCTTCGTATACATGTACGCTCA GTCANTTTTCGGTTATATACCGTCACTTGCAAACTGCGAGTTCGC
fabH_mut2	RBS degeneracy	C*A*G*A*TAGCTGCCAGTACCAATAATCTTCGTATACATGTACGCTCA GTNACTTTTCGGTTATATACCGTCACTTGCAAACTGCGAGTTCGC
fabH_mut3	RBS degeneracy	C*A*G*A*TAGCTGCCAGTACCAATAATCTTCGTATACATGTACGCTCA GNCACTTTTCGGTTATATACCGTCACTTGCAAACTGCGAGTTCGC
fabH_mut4	RBS degeneracy	C*A*G*A*TAGCTGCCAGTACCAATAATCTTCGTATACATGTACGCTNA GTCACTTTTCGGTTATATACCGTCACTTGCAAACTGCGAGTTCGC
fabH_mut5	RBS degeneracy	C*A*G*A*TAGCTGCCAGTACCAATAATCTTCGTATACATGTACGCNCA GTCACTTTTCGGTTATATACCGTCACTTGCAAACTGCGAGTTCGC
fabH_mut6	RBS degeneracy	C*A*G*A*TAGCTGCCAGTACCAATAATCTTCGTATACATGTACGNTCA GTCACTTTTCGGTTATATACCGTCACTTGCAAACTGCGAGTTCGC

Supplementary Table S2.5.
Oligonucleotides used for glucaric acid mutagenesis. * symbol indicates phosphortioate bonds.

Name	Locus	Function	Sequence
garKnull	*garK*	stop	T*T*C*C*CGAAATCCTTTTTCTATCGCCTGCGCAACCTCGCTGGCAGATTAA CTTTCTTTTTAAGAGTCTGGGGCGATTACGATTTTCATACC
uxaCnull	*uxaC*	stop	C*A*A*A*TGGCAATGGTAATCGAAAATCGGCTGGTCTTTTGCTTAGTCGTGT TACAGACGGCGGGCAAATTCGGTATCTAACAGGAAATCTTC
suhB-degen	*suhB*	RBS degeneracy	C*T*C*G*CTGCTATACTCTGCGCCGTTTTCCCGTTCTTTAACATCCDDVVVV VDDDDCCGATGCATCCGATGCTGAACATCGCCGTGCGCGCA
pgi-degen	*pgi*	RBS degeneracy	G*G*C*A*GCGGTCTGCGTTGGATTGATGTTTTTCATTAGHHHHBBBBBHHTG ATTTTGAGAATTGTGACTTTGGAAGATTGTAGCGCCAGTCA
sthA-degen	*sthA*	RBS degeneracy	C*G*C*G*ATAAAATGTTACCATTCTGTTGCTTTTATGTATAAGAACDDVVVV VDDDDACCATGCCACATTCCTACGATTACGATGCCATAGTA
zwf-degen	*zwf*	RBS degeneracy	G*T*C*A*CAGGCCTGGGCTGTTTGCGTTACCGCCATGTCHHHHBBBBBHHGT TAACTAACCCGGTACTTAAGCCAGGGTATACTTGTAATTTT
mdh-degen	*mdh*	RBS degeneracy	A*C*C*G*CCAGCAGCGCCGAGGACTGCGACTTTCATCCTHHHHBBBBBHHTA TATTGATAAACTAAGATATGTTGCTCCGCTGCCGCGACCTT

Supplementary Table S2.6
List of key candidate genes predicted by flux balance analysis to increase naringenin production.

Reaction	Expression prediction	Gene EcoCyc B-number(s)
ACCOAC	Up	b0185 and b2316 and b3255 and b3256
G6PDH2r	Down	b1852
GAPD	Up	b1779
CS	Down	b0720
FUM	Down	b1612 or b4122 or b1611
TKT2	Down	b2935
RPE	Down	b3386
PGM	Up	b0755
ENO	Up	b2779
F6PA	Up	b0825
PGK	Up	b2926
DHAPT	Up	b1200 and b1199 and b1198 and b2415 and b2416
TPI	Up	b3919
TALA	Up	b0008
PYK	Up	b1676
PGI	Up	b4025
THD2pp	Up	b1602 and b1603
ACONT	Down	b0118
MDH	Down	b3236
PDH	Up	b0114 and b0115 and b0116
TALA	Up	b0008

Supplementary Table S2.7
Genomic mutations found in evolved production strains. Genomic loci reference the *E. coli* MG1655 genome position. Types of mutations refer to the following changes: RBS, ribosome binding site mutation; start codon, alternate translation start codon mutation; coding, missense mutation; frameshift, insertion or deletion leading to a frameshifted reading frame; nonsense, premature stop codon mutation.

Supplementary Table 2.7 (Continued)

Naringenin production strains

Strain	Gene name	Type of change	Genomic locus	Mutation
Strain 1	*fabD*	RBS	1148940	A->G
	fabF	Start codon	1151162	GTG->TTG
	fabH	Start codon	1147982	ATG->GTG
	fumC	Start codon	1684612	ATG->TTG
	entB	Frameshift	627205	G insertion
	rpoD	RBS	3211052	A->G
Strain 2	*fabF*	Start codon	1151162	GTG->TTG
	fabH	Start codon	1147982	ATG->GTG
	fumC	Start codon	1684612	ATG->TTG
	entB	Frameshift	627205	G insertion
	rpoD	RBS	3211052	A->G
Strain 3	*fabD*	RBS	1148939	A->T
	fabF	Start codon	1151162	GTG->TTG
	fabH	Start codon	1147982	ATG->GTG
	fumC	Start codon	1684612	ATG->TTG
	entB	Frameshift	627205	G insertion
	rpoD	RBS	3211052	A->G
Strain 4	*fabD*	RBS	1148940	A->G
	fabH	Start codon	1147982	ATG->GTG
	fumC	Start codon	1684612	ATG->TTG
	aroG	Coding	785291	G->A
	entB	Frameshift	627205	G insertion
	rpoD	RBS	3211052	A->G
Strain 5	*fabF*	RBS	1151156	A->C
	fabH	Start codon	1147982	ATG->GTG
	fumC	Start codon	1684612	ATG->TTG
	hcaT	Frameshift	2665603	C insertion
	entB	Frameshift	627205	G insertion
	rpoD	RBS	3211052	A->G
Strain 6	*fabF*	Start codon	1151162	GTG->TTG
	fabH	Start codon	1147982	ATG->GTG
	fumC	Start codon	1684612	ATG->TTG
	tyrA	Coding	2737031	G->A
	entB	Frameshift	627205	G insertion
	rpoD	RBS	3211052	A->G
Strain 7	*fabD*	RBS	1148942	G->C
	fabH	RBS	1149790	G->C
	fumC	Start codon	1684612	ATG->TTG
	tyrA	Coding	2737031	G->A
	mhpD	Frameshift	372032	G deletion
	entB	Frameshift	627205	G insertion
	rpoD	RBS	3211052	A->G

Glucaric acid production strain

Strain	Gene name	Type of change	Genomic locus	Mutation
Strain 1	*garK*	Nonsense	3269752	A->T
	garK	Nonsense	3269752	A->T

Supplementary Table S2.8

Number of mutations found in evolved pathway strains.

Naringenin production strains

Strain name	Start codon	Frameshift/Nonsense	Non-synonymous
Strain 1	3	10	65
Strain 2	3	9	67
Strain 3	3	10	67
Strain 4	2	10	70
Strain 5	2	11	65
Strain 6	3	9	70
Strain 7	1	11	67

Glucaric acid production strain

Strain name	Start codon	Frameshift/Nonsense	Non-synonymous
Strain 1	2	8	32

Supplementary Tables for Chapter 3

Supplementary Table S3.1

LacI mutants responsive to new inducers. Induction values represent FACS fold induction measurement for given inducers supplemented into induction medium at 3 mM.

Fucose-responsive variants	
Fucose induction	Mutant
10.48	Q291T
9.59	I289V, Q291T, L330N
7.05	I289V, Q291T, L296I
5.17	I79Q
4.66	Q291I
4.48	A245G, I289V, Q291S, L296I
4.04	Q291S
3.89	I79A
3.66	I79M
3.57	Q291A
3.44	I79L
3.09	Y273F, Q291V, L296I
2.98	Y273H, Q291H
2.6	I79S
2.57	I79L, D130E
2.52	L73N, I79L, S97T
2.35	I79S
2.27	I79F
2.27	Q291E
2.21	L73N, I79S
2.1	L73N
Lactitol-responsive variants	
Lactitol induction	Mutant
8.99	I79G
6.26	I79S
5	I79A
4.36	I79T, R101H
3.72	I79T, V104Q
3.7	I79Y
3.59	I79Y, K108I
3.44	I79T
3.42	I79Y, R101Q
3.22	I79C, G103S
3.13	I79T, G103D
3.07	I79F
2.93	I79S, K108R
2.59	I79T, G91M
2.51	I79C, V94L
2.3	I79Q
2.26	D149G

Supplementary Table S3.1 (Continued)

Lactitol induction	Mutant
2.12	I79T, Q89A
2.11	I79L
Sucralose-responsive variants	
Sucralose induction	Mutant
11.3	D149T, V150A, I156L, S193D
10.35	D149G, I160V, H163Y, S193E
8.46	L148N, I160N, H163W, S191D, S193N
7.2	I160S, H163W, S191A, L196R
6.62	D149G, I160V, H163W, S193A
4.95	I160N, L196R
4.85	I160N, H163D, S193A, L196R
4.77	I160N, H163Y, L196R
4.52	N246D,Y273H
4.5	I160Q, H163Y, L196R
3.95	I160D, H163W, L196R
3.43	Q291I, F293R
3.27	I79A
3.05	D149G, I160T, H163Y, S193Q
2.65	A43V, I79A
2.64	Y273H, Q291I, F293W
2.63	L196K
2.48	Y273H, Q291L, L296V
2.44	L148T, I160V, F161S, H163Y, S193A
2.37	P76A, N125S
2.36	Q291V, L296I
2.35	D149G, I160N, H163Y, E164G, S193A
2.35	I79A, N125S
2.27	L73A, A75G, P76A, I79A, N125S
2.24	P76A, I79V, N125D
2.15	I79S
2.14	N246T, Q291V
2.06	L73K, I79A
2.04	L73Y, I79V
2	L73Q, I79V, K84E
2	Y273F, Q291E
Gentiobiose-responsive variants	
Gentiobiose induction	Mutant
10.17	H173Y, Q291H
9.77	S70D, H74S
8.87	I213P, Q291H
8.55	R207H, Q209T
8.51	R255H, S269A
8.36	R255H
8.05	T5S, A32Q
7.84	L63P, H74S
7.76	Q291H

Supplementary Table S3.1 (Continued)

Gentiobiose induction	Mutant
7.5	S69Y, H74S
7.33	R255C, P285A
6.68	V20A
6.59	Q227G
6.4	G166S, A176T, G236H
6.21	S221D
6.1	Q227I
5.99	F226C
5.98	N25G
5.98	S221Q
5.79	T34L
5.19	R255H, C281M
5.17	M223S
5.17	Y126T, T141A
5.16	A250C
4.98	S221C, G236C
4.7	Q228Y
4.65	L115S
4.62	H112D, A199T
4.4	D278A
4.4	H112G, D130Y
4.35	S280W
4.22	G297A, K314R
4.22	R294D
4.15	K327R
4.14	V52Q, H74S
4.12	S224M, E235D
4.05	V15I
3.91	G297A
3.84	S221C, T229A
3.83	D219L
3.81	H29E
3.78	S61E, H74S
3.77	P285A
3.64	V244I
3.61	T288C
3.6	S21A
3.56	M249Q
3.54	M232I, V324Q
3.53	F226S, Q231A
3.49	R255C
3.39	V238R
3.36	S97H
3.26	R263M
3.23	T34C
3.2	L71K, H74S

Supplementary Table S3.1 (Continued)

Lactitol induction	Mutant
3.18	E277D, C281F
3.09	A199T, S221C, Q228I
3.08	L356G
3.07	S280Y
3.01	R101W
2.99	I283N
2.98	R255C, P320C
2.98	V30I, L177D
2.92	S85A
2.9	L342C
2.78	F226R, Q227R
2.76	L115S, D130G, T141I
2.75	V24K
2.73	Q89D
2.72	A27E
2.71	S120G, V143N
2.67	D8Q, T34M
2.67	V99D
2.6	G297S
2.59	A199K
2.53	K33D
2.52	L114F, T141E
2.52	T229V
2.51	N25R
2.5	Q227W, E235D
2.48	V94C, R101L
2.46	V104S, A110V
2.46	V99I, R101Y
2.43	C281M
2.39	A110C
2.39	L63N, H74S
2.39	S85D
2.36	P239W, M254I
2.33	D278I
2.33	L6I, G14D
2.24	S31G, A32V
2.23	G310Y
2.23	R294L
2.21	A245C, Q248H
2.21	S70N, L71M, H74S
2.17	A350C
2.13	A116P, T141C
2.12	N46H, H74S
2.1	Q227R, M242Q
2.09	V238E, P239S
2.08	A214Y

Supplementary Table S3.1 (Continued)

Lactitol induction	Mutant
2.06	S85G, V99C, G103S
2.04	A199C, A343V
2.04	A27P
2.04	R35F
2.03	P339R
2.02	Q211R, R294L, V313A

Supplementary Table S3.2

Comparison of success rate of computational protein design and protein-wide single amino acid substitution vs. error prone PCR.

Maximum fold induction variant comparisons			
	Fucose	Lactitol	Sucralose
Computational protein design	10.5	7.1	11.3
Error-prone PCR	5.0	4.8	1.7
Percent of clones that are new variants with fold induction greater than 2.0			
	Fucose	Lactitol	Sucralose
Computational protein design	42.7	27.1	41.6
Error-prone PCR	17.7	5.2	0.0

Maximum fold induction variant comparison	
	Gentiobiose
Error-prone PCR	9.1
Single amino acid substitution	8.9
Percent of clones that are new variants with fold induction greater than 2.0	
	Gentiobiose
Error-prone PCR	17.7
Single amino acid substitution	19.8

Supplementary Table S3.3

Mutations conferring I^S phenotype to LacI. Forty-four LacI variants not responsive to IPTG (I^S) used for shuffling to obtain enhanced specificity.

H74Y	N246T
L73S, H74S	Q227G, Q248R
S61A, H74Q, A75P	Q248G
S69E, H74Q, A75R	R197C, A250C
S70H, H74Q, A75P	V238L, Q248H
S70N, H74F, A75T	V244I, Q248R
H74Y, S102C	A266L, T276I
D149F	D274E
D149M	D274G
D149Q, S151P, A176T	D274H
D149W	D274T
D149Y, Q180R	D275G, T276F
H74Y, A176F	D275G
I159H	D275L
I159L	T276A
S93N, D149F	T276F
A245S, Q248P	T276K
H202Y, Q248G	T276L, S279G
I237T, Q248R	T276N
N246G, Q248H, A253T	T276V
N246K	T276W
N246Q	F293Y

Supplementary Table S3.4

X-ray data collection and refinement statistics.

[a] $R_{merge}(I) = \sum_{hkl} ((\sum_i |I_{hkl,i} - \langle I_{hkl} \rangle|)/\sum_i I_{hkl,i})$

[b] $R_{work} = \sum_{hkl} |F_{obs} - F_{calc}|/\sum_{hkl} F_{obs}$

[c] $R_{free} = \sum_{hkl} |F_{obs} - F_{calc}|/\sum_{hkl} F_{obs}$, where all reflections belong to a test set of 5% randomly selected data

Active-Site Ligand	***Unliganded***		***Sucralose-bound***	
PDB accession ID	**4RZS**		**4RZT**	
Data collection statistics				
Resolution range, overall (Å)	94.49 – **2.71**		81.67 – **3.03**	
Resolution range, last shell (Å)	2.78 – 2.71		3.15 – 3.03	
Space group	$P2_12_12_1$		C2	
Unit cell dimensions				
a (Å)	91.1		163.6	
b (Å)	111.6		74.1	
c (Å)	189.3		149.2	
α (°)	90.0		90.0	
β (°)	90.0		119.9	
γ (°)	90.0		90.0	
Radiation source	APS 24-ID-C		APS 24-ID-C	
Radiation wavelength (Å)	0.9793		0.9791	
	Statistics in parentheses were obtained after ellipsoidal truncation by the Diffraction Anisotropy Server			
Diffraction Data Processing Statistics				
Anisotropic truncation high resolution cutoff range (Å)	**3.3** to **2.7**		**3.5** to **3.1**	
Measured reflections	349,936	(233,832)	95,460	(81,944)
Unique reflections	53,814	(36,390)	25,828	(24,038)
Overall completeness (%)	99.6	(68.2)	97.8	(84.8)
Last shell completeness (%)	97.9	(3.1)	89.2	(11.3)
Overall R_{merge} [a]	0.086	(0.069)	0.088	(0.079)
Last shell R_{merge}	9.042	(0.461)	1.315	(58.0)
Overall I/σ(*I*)	10.5	(15.2)	10.6	(12.2)
Last shell I/σ(*I*)	0.4	(3.9)	0.9	(2.0)
Refinement Statistics				
R_{work} [b]	0.199		0.194	
R_{free} [c]	0.242		0.231	
R_{free} test set size (%)	5 (random)		5 (random)	
rmsd bond length (Å)	0.01		0.01	
rmsd bond angle (°)	1.54		1.28	
Structure Validation Statistics				
Ramachandran plot				
Most favored (%)	87.5		87.8	
Additional allowed regions (%)	10.2		9.7	
Generously allowed regions (%)	1.4		1.7	
Disallowed regions (%)	1.0		0.8	

Supplementary Table S3.5

LacI gene and expression plasmid sequences

pSC101_LacI_specR plasmid sequence
CAATTCCGACGTCTAAGAAACCATTATTATCATGACATTAACCTATAAAAATAGGCGTATCACGAGGCCCTTTCGTC
TTCACCTCGAGTCCCTATCAGTGATAGAGATTGACATCCCTATCAGTGATAGAGATACTGAGCACATCAGCAGGACG
CACTGACCGAATTCATTAAAGAGGAGAAAGGTACCGTGAAACCAGTAACGTTATACGATGTCGCAGAGTATGCCGGT
GTCTCTTATCAGACCGTTTCCCGCGTGGTGAACCAGGCCAGCCACGTTTCTGCGAAAACGCGGGAAAAAGTGGAAGC
GGCGATGGCGGAGCTGAATTACATTCCCAACCGCGTGGCACAACAACTGGCGGGCAAACAGTCGTTGCTGATTGGCG
TTGCCACCTCCAGTCTGGCGCTGCACGCGCCGTCGCAAATTGTCGCGGCGATTAAATCTCGCGCCGATCAACTGGGT
GCCAGCGTGGTGGTGTCGATGGTAGAACGAAGCGGCGTCGAAGCCTGTAAAGCGGCGGTGCACAATCTTCTCGCGCA
ACGCGTCAGTGGGCTGATCATTAACTATCCGCTGGATGACCAGGATGCCATTGCTGTGGAAGCTGCCTGCACTAATG
TTCCGGCGTTGTTTCTTGATGTCTCTGACCAGACACCCATCAACAGTATTATTTTCTCCCATGAAGACGGTACGCGA
CTGGGCGTGGAGCATCTGGTCGCATTGGGTCACCAGCAAATCGCGCTGTTAGCGGGCCCATTAAGTTCTGTCTCGGC
GCGTCTGCGTCTGGCTGGCTGGCATAAATATCTCACTCGCAATCAAATTCAGCCGATAGCGGAACGGGAAGGCGACT
GGAGTGCCATGTCCGGTTTTCAACAAACCATGCAAATGCTGAATGAGGGCATCGTTCCCACTGCGATGCTGGTTGCC
AACGATCAGATGGCGCTGGGCGCAATGCGCGCCATTACCGAGTCCGGGCTGCGCGTTGGTGCGGATATCTCGGTAGT
GGGATACGACGATACCGAAGACAGCTCATGTTATATCCCGCCGTTAACCACCATCAAACAGGATTTTCGCCTGCTGG
GGCAAACCAGCGTGGACCGCTTGCTGCAACTCTCTCAGGGCCAGGCGGTGAAGGGCAATCAGCTGTTGCCCGTCTCA
CTGGTGAAAAGAAAAACCACCCTGGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAATGCA
GCTGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAGTGAATTAGCAGAAAGTCAAAAGCCTCCGACCGGAGGCTTT
TGACTAAAACTTCCCTTGGGGTTATCATTGGGGCTCACTCAAAGGCGGTAATCAGATAAAAAAAATCCTTAGCTTTC
GCTAAGGATGATTTCTGCTAGTATTATTATTTGCCGACTACCTTGGTGATCTCGCCTTTCACGTAGTGGACAAATTC
TTCCAACTGATCTGCGCGCGAGGCCAAGCGATCTTCTTCTTGTCCAAGATAAGCCTGTCTAGCTTCAAGTATGACGG
GCTGATACTGGGCCGGCAGGCGCTCCATTGCCCAGTCGGCAGCGACATCCTTCGGCGCGATTTTGCCGGTTACTGCG
CTGTACCAAATGCGGGACAACGTAAGCACTACATTTCGCTCATCGCCAGCCCAGTCGGGCGGCGAGTTCCATAGCGT
TAAGGTTTCATTTAGCGCCTCAAATAGATCCTGTTCAGGAACCGGATCAAAGAGTTCCTCCGCCGCTGGACCTACCA
AGGCAACGCTATGTTCTCTTGCTTTTGTCAGCAAGATAGCCAGATCAATGTCGATCGTGGCTGGCTCGAAGATACCT
GCAAGAATGTCATTGCGCTGCCATTCTCCAAATTGCAGTTCGCGCTTAGCTGGATAACGCCACGGAATGATGTCGTC
GTGCACAACAATGGTGACTTCTACAGCGCGGAGAATCTCGCTCTCTCCAGGGGAAGCCGAAGTTTCCAAAAGGTCGT
TGATCAAAGCTCGCCGCGTTGTTTCATCAAGCCTTACGGTCACCGTAACCAGCAAATCAATATCACTGTGTGGCTTC
AGGCCGCCATCCACTGCGGAGCCGTACAAATGTACGGCCAGCAACGTCGGTTCGAGATGGCGCTCGATGACGCCAAC
TACCTCTGATAGTTGAGTCGATACTTCGGCGATCACCGCTTCCCTCATGATGTTTAACTTTGTTTTAGGGCGACTGC
CCTGCTGCGTAACATCGTTGCTGCTCCATAACATCAAACATCGACCCACGGCGTAACGCGCTTGCTGCTTGGATGCC
CGAGGCATAGACTGTACCCCAAAAAAACAGTCATAACAAGCCATGAAAACCGCCACTGCGCCGTTACCACCGCTGCG
TTCGGTCAAGGTTCTGGACCAGTTGCGTGAGCGCATAACACCCCTTGTATTACTGTTTATGTAAGCAGACAGTTTTA
TTGTTCATGATGATATATTTTTATCTTGTGCAATGTAACATCAGAGATTTTGAGACACAACGTGGCTTTGTTGAATA
AATCGAACTTTTGCTGAGTTGAAGGATCAGGTTACATTGTCGATCTGTTCATGGTGAACAGCTTTAAATGCACCAAA
AACTCGTAAAAGCTCTGATGTATCTATCTTTTTTACACCGTTTTCATCTGTGCATATGGACAGTTTTCCCTTTGATA
TCTAACGGTGAACAGTTGTTCTACTTTTGTTTGTTAGTCTTGATGCTTCACTGATAGATACAAGAGCCATAAGAACC
TCAGATCCTTCCGTATTTAGCCAGTATGTTCTCTAGTGTGGTTCGTTGTTTTTGCGTGAGCCATGAGAACGAACCAT
TGAGATCATGCTTACTTTGCATGTCACTCAAAAATTTTGCCTCAAAACTGGTGAGCTGAATTTTTGCAGTTAAAGCA
TCGTGTAGTGTTTTTCTTAGTCCGTTACGTAGGTAGGAATCTGATGTAATGGTTGTTGGTATTTTGTCACCATTCAT
TTTTATCTGGTTGTTCTCAAGTTCGGTTACGAGATCCATTTGTCTATCTAGTTCAACTTGGAAAATCAACGTATCAG
TCGGGCGGCCTCGCTTATCAACCACCAATTTCATATTGCTGTAAGTGTTTAAATCTTTACTTATTGGTTTCAAAACC
CATTGGTTAAGCCTTTTAAACTCATGGTAGTTATTTTCAAGCATTAACATGAACTTAAATTCATCAAGGCTAATCTC
TATATTTGCCTTGTGAGTTTTCTTTTGTGTTAGTTCTTTTAATAACCACTCATAAATCCTCATAGAGTATTTGTTTT
CAAAAGACTTAACATGTTCCAGATTATATTTTATGAATTTTTTTAACTGGAAAAGATAAGGCAATATCTCTTCACTA
AAAACTAATTCTAATTTTTCGCTTGAGAACTTGGCATAGTTTGTCCACTGGAAAATCTCAAAGCCTTTAACCAAAGG
ATTCCTGATTTCCACAGTTCTCGTCATCAGCTCTCTGGTTGCTTTAGCTAATACACCATAAGCATTTTCCCTACTGA
TGTTCATCATCTGAGCGTATTGGTTATAAGTGAACGATACCGTCCGTTCTTTCCTTGTAGGGTTTTCAATCGTGGGG
TTGAGTAGTGCCACACAGCATAAAATTAGCTTGGTTTCATGCTCCGTTAAGTCATAGCGACTAATCGCTAGTTCATT
TGCTTTGAAAACAACTAATTCAGACATACATCTCAATTGGTCTAGGTGATTTTAATCACTATACCAATTGAGATGGG
CTAGTCAATGATAATTACATGTCCTTTTCCTTTGAGTTGTGGGTATCTGTAAATTCTGCTAGACCTTTGCTGGAAAA
CTTGTAAATTCTGCTAGACCCTCTGTAAATTCCGCTAGACCTTTGTGTGTTTTTTTTGTTTATATTCAAGTGGTTAT

Supplementary Table S3.5 (Continued)

pSC101_LacI_specR plasmid sequence (continued)
AATTTATAGAATAAAGAAAGAATAAAAAAAGATAAAAAGAATAGATCCCAGCCCTGTGTATAACTCACTACTTTAGT CAGTTCCGCAGTATTACAAAAGGATGTCGCAAACGCTGTTTGCTCCTCTACAAAACAGACCTTAAAACCCTAAAGGC TTAAGTAGCACCCTCGCAAGCTCGGGCAAATCGCTGAATATTCCTTTTGTCTCCGACCATCAGGCACCTGAGTCGCT GTCTTTTTCGTGACATTCAGTTCGCTGCGCTCACGGCTCTGGCAGTGAATGGGGGTAAATGGCACTACAGGCGCCTT TTATGGATTCATGCAAGGAAACTACCCATAATACAAGAAAAGCCCGTCACGGGCTTCTCAGGGCGTTTTATGGCGGG TCTGCTATGTGGTGCTATCTGACTTTTTGCTGTTCAGCAGTTCCTGCCCTCTGATTTTCCAGTCTGACCACTTCGGA TTATCCCGTGACAGGTCATTCAGACTGGCTAATGCACCCAGTAAGGCAGCGGTATCATCAACAGGCTTACCCGTCTT ACTGTCCCTAGTGCTTGGATTCTCACCAATAAAAAACGCCCGGCGGCAACCGAGCGTTCTGAACAAATCCAGATGGA GTTCTGAGGTCATTACTGGATCTATCAACAGGAGTCCAAGCGAGCTCGTAAACTTGGTCTGACAGCTCTAGCTCCGG CAAAAAAACGGGCAAGGTGTCACCACCCTGCCCTTTTTCTTTAAAACCGAAAAGATTACTTCGCGTTTGCCACCTGA CGTCTAAGAAAAGGAATATTCAGCAATTTGCCCGTGCCGAAGAAAGGCCCACCCGTGAAGGTGAGCCAGTGAGTTGA TTGCTACGTAA
lacI recoded gene sequence
GTGAAACCAGTAACGTTATACGATGTCGCAGAGTATGCCGGTGTCTCTTATCAGACCGTTTCCCGCGTGGTGAACCA GGCCAGCCACGTTTCTGCGAAAACGCGGGAAAAAGTGGAAGCGGCGATGGCGGAGCTGAATTACATTCCCAACCGCG TGGCACAACAACTGGCGGGCAAACAGTCGTTGCTGATTGGCGTTGCCACCTCCAGTCTGGCGCTGCACGCGCCGTCG CAAATTGTCGCGGCGATTAAATCTCGCGCCGATCAACTGGGTGCCAGCGTGGTGGTGTCGATGGTAGAACGAAGCGG CGTCGAAGCCTGTAAAGCGGCGGTGCACAATCTTCTCGCGCAACGCGTCAGTGGGCTGATCATTAACTATCCGCTGG ATGACCAGGATGCCATTGCTGTGGAAGCTGCCTGCACTAATGTTCCGGCGTTGTTTCTTGATGTCTCTGACCAGACA CCCATCAACAGTATTATTTTCTCCCATGAAGACGGTACGCGACTGGGCGTGGAGCATCTGGTCGCATTGGGTCACCA GCAAATCGCGCTGTTAGCGGGCCCATTAAGTTCTGTCTCGGCGCGTCTGCGTCTGGCTGGCTGGCATAAATATCTCA CTCGCAATCAAATTCAGCCGATAGCGGAACGGGAAGGCGACTGGAGTGCCATGTCCGGTTTTCAACAAACCATGCAA ATGCTGAATGAGGGCATCGTTCCCACTGCGATGCTGGTTGCCAACGATCAGATGGCGCTGGGCGCAATGCGCGCCAT TACCGAGTCCGGGCTGCGCGTTGGTGCGGATATCTCGGTAGTGGGATACGACGATACCGAAGACAGCTCATGTTATA TCCCGCCGTTAACCACCATCAAACAGGATTTTCGCCTGCTGGGGCAAACCAGCGTGGACCGCTTGCTGCAACTCTCT CAGGGCCAGGCGGTGAAGGGCAATCAGCTGTTGCCCGTCTCACTGGTGAAAAGAAAAACCACCCTGGCGCCCAATAC GCAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAATGCAGCTGGCACGACAGGTTTCCCGACTGGAAAGCGGGC AGTGA

Supplementary Table S3.6
Primers used in this study

Backbone amplification primers for subpool cloning	
Primer name	Sequence
backbone_LacI-F_tile1	CGGTAGCGCATATGGGGTCTCAGAAGCGGCGATGGCGGA
backbone_LacI-R_tile1	ACAGGACGTTCGGGTCTCCTCACGGTACCTTTCTCCTCTTTAATG
backbone_LacI-F_tile2	CGCTCTGTTTGACGGGGTCTCGCACTTTTTCCCGCGTTTTCGCA
backbone_LacI-R_tile2	CCACACCTGTAAGGTGGTCTCTTCCGTCGCAAATTGTCGC
backbone_LacI-F_tile3	CGAACGGCAACCCTCGGTCTCAGCAGGGCCAGACTGGAG
backbone_LacI-R_tile3	CGGAGAGATAACTACGGTCTCCGTTCACAATCTTCTCGCGCAAC
backbone_LacI-F_tile4	AGTGCCGCTTACAGCGGTCTCGTTCTTGATGTCTCTGACCAGACA
backbone_LacI-R_tile4	GACGTCTGTAGTCTGGGTCTCTGCTGCTTTACAGGCTTCGAC
backbone_LacI-F_tile5	GACGTCTGTAGTCTGGGTCTCTGCTGTTAGCGGGACCATTA
backbone_LacI-R_tile5	CACGCTATTGAGGCAGGTCTCCTAACGCCGGAACATTAGTGCAG
backbone_LacI-F_tile6	GGACATCCGGAATGAGGTCTCGGACTGGAGTGCCATGTCC
backbone_LacI-R_tile6	ACCGGATCTCCTTAGGGTCTCTGATTTGCTGGTGACCCAATG
backbone_LacI-F_tile7	CGGTAGCGCATATGGGGTCTCATGCCATTACCGAGTCCG
backbone_LacI-R_tile7	TCCACAGGACGTTCGGGTCTCCCTTCCCGTTCCGCTATCG
backbone_LacI-F_tile8	CGCTCTGTTTGACGGGGTCTCGAGGATTTTCGCCTGCTGG
backbone_LacI-R_tile8	CCACACCTGTAAGGTGGTCTCTTTGCGCCCAGCGCCAT
backbone_LacI-F_tile9	GACGTCTGTAGTCTGGGTCTCTAACCACCCTGGCACCCAAT
backbone_LacI-R_tile9	CGGAGAGATAACTACGGTCTCCTGATGGTGGTTAACGGC
backbone_LacI-F_tile10	AGTGCCGCTTACAGCGGTCTCGCAGTGAAGCTTAGCAGAAAGTC
backbone_LacI-R_tile10	GACGTCTGTAGTCTGGGTCTCTTCTTTTCACCAGTGAGACGG
backbone_LacI-F_73-125	GGACATCCGGAATGAGGTCTCGTATCCGCTGGATGACCAGGATGCCA
backbone_LacI-R_73-125	ACCGGATCTCCTTAGGGTCTCTCGCCAGACTGGAGGTGGCAACGCCA
backbone_LacI-F_148-197	CGGTAGCGCATATGGGGTCTCACTGGCTGGCTGGCATAAATATCTCACTCGCA
backbone_LacI-R_148-197	TCCACAGGACGTTCGGGTCTCCAAACAACGCCGGAACATTAGTGCAGGCA
backbone_LacI-F_245-296	CGCTCTGTTTGACGGGGTCTCGGGGCAAACCAGCGTGGACCGCTT
backbone_LacI-R_245-296	CCACACCTGTAAGGTGGTCTCTAACCAGCATCGCAGTGGGAACGATG
Subpool primers for orthongonal amplification	
Primer name	Sequence
skpp15-1-F	GGGTCACGCGTAGGA
skpp15-2-F	CGCGTCGAGTAGGGT
skpp15-3-F	CGATCGCCCTTGGTG
skpp15-4-F	GGTCGAGCCGGAACT
skpp15-5-F	TCCCGGCGTTGTCCT

Supplementary Table S3.6 (Continued)

skpp15-6-F	CGCAGGGTCCAGAGT
skpp15-7-F	AGTGACCCGTCCCTG
skpp15-8-F	TGCCCGCTGTCTTCA
skpp15-9-F	CGATCGTGCCCACCT
skpp15-10-F	ACTGGTGCGTCGTCT
skpp15-11-F	AGCGAAACCGTGCGT
skpp15-12-F	ACCGGTTTCCACGCA
skpp15-13-F	GGGTTCGAGCGGGAG
skpp15-14-F	TGGGCGCCAAGAACC
skpp15-15-F	ACTCGACGGCCTCTG
skpp15-16-F	CCCGGATCCCTTGCT
skpp15-17-F	GCTTCCGCCCGTAGA
skpp15-18-F	CGAGCCGTGGTTCCT
skpp15-19-F	ACGCCGAATCCCACA
skpp15-20-F	ATCACTCGCGTCCCA
skpp15-21-F	ACCATCGCGCACCTT
skpp15-22-F	CGTCACGCAGGGTTC
skpp15-23-F	AGCTGCTACACCGCC
skpp15-24-F	ATGGACGCGTGGAGT
skpp15-25-F	AGTTGACCAGCGCCA
skpp15-26-F	GCGGCACCACAAACT
skpp15-27-F	ACCTTCACGCGTCCC
skpp15-28-F	GACTGCGGCGTTGGT
skpp15-29-F	GCCGCCGCTTATTGG
skpp15-30-F	TCCACCGTCGGCAAG
skpp15-31-F	GGCCGGACACATGCT
skpp15-32-F	ACGGGCTGGTGTGGT
skpp15-33-F	AAGTGCCCTTCCCGT
skpp15-34-F	CGCGCCGTTCCTTTG
skpp15-35-F	TGCGGGCCCATTTCC
skpp15-36-F	CGAAGTGCGCAGGGT
skpp15-1-R	GTTCCGCAGCCACAC
skpp15-2-R	GCCGTGTGAAGCTGG
skpp15-3-R	GGTTTAGCCGGCGTG
skpp15-4-R	GGATGCGCACCCAGA
skpp15-5-R	GCTCCGTCACTGCCC
skpp15-6-R	GTTCGCGCGAAGGAA
skpp15-7-R	AGTCGACCTCTGCCC

Supplementary Table S3.6 (Continued)

skpp15-8-R	CACTGCCGGGCTACA
skpp15-9-R	GTGCGGGCTCCAACT
skpp15-10-R	CGGGAAGTGTTCGCC
skpp15-11-R	ACTCTCCCGCGTCCC
skpp15-12-R	CCTCCAAGCGCGAAC
skpp15-13-R	TAGCGCGCAGAGAGG
skpp15-14-R	CCCACTGCGCCCAAG
skpp15-15-R	ACACGCGCGTTGAAG
skpp15-16-R	CGCGTCATTGGCCTC
skpp15-17-R	CATTCGCGCTGGTGG
skpp15-18-R	CGCTGGGAGGGTGTT
skpp15-19-R	TGCTGGAACGTCGGG
skpp15-20-R	ACACGCTCGACTCCC
skpp15-21-R	CTCCCTTGCCTGCCC
skpp15-22-R	ATCCCGCCGTCAGAG
skpp15-23-R	GCGCGATGGTCACAG
skpp15-24-R	TAGGGCTCCGCTTGG
skpp15-25-R	AGTCGCGCCTACCAC
skpp15-26-R	CGTGGCCTCTGTCCT
skpp15-27-R	GCCCACCGACTCCAC
skpp15-28-R	TACGCCCGGGACAGA
skpp15-29-R	TGCGCCACCTCAGTC
skpp15-30-R	GGCCGCACCCAGTAG
skpp15-31-R	CGGTGTGGCTGACGG
skpp15-32-R	CTGGGTGGCGCAGTC
skpp15-33-R	GAGTCCGCGCAAGAG
skpp15-34-R	TCGTCGCGCAGGTTC
skpp15-35-R	GCCCACCGTCGCTTC
skpp15-36-R	TTGCGCCCGTGTCAG
Primers for error-prone PCR cloning	
Primer name	Sequence
LacI_EP_5primeF	ACGGCAACCCTCGGTCTCAGCAAACAGTCGTTGCTGATTGGCGTT
LacI_EP_3primeR	CTCTGTTTGACGGGGTCTCAGCAGCAAGCGGTCCACGCTGGTTTG
BB_EP_3primeF	CCACACCTGTAAGGTGGTCTCACTGCAACTCTCTCAGGGCCAGGC
BB_EP_5primeR	CGGTAGCTCATATGGGGTCTCATTGCCCGCCAGTTGTTGTGCCAC
Primers used for MiSeq library preparation – index sequences underlined	
Primer name	Sequence
LacI_MiSeq_1F	CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNGAGCACATCAGCAGGACG

Supplementary Table S3.6 (Continued)

LacI_MiSeq_1R	GGAGTTCAGACGTGTGCTCTTCCGATCTTCGCCGCGACAATTTGCGA
LacI_MiSeq_2F	CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNCGTTGCCACCTCCAGTCTG
LacI_MiSeq_2R	GGAGTTCAGACGTGTGCTCTTCCGATCTTTCATGGGAGAAAATAATACTGTTGATG
LacI_MiSeq_3F	CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNAGCTGCCTGCACTAATGTTC
LacI_MiSeq_3R	GGAGTTCAGACGTGTGCTCTTCCGATCTCTCATTCAGCATTTGCATGGT
LacI_MiSeq_4F	CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNCAGCCGATAGCGGAACG
LacI_MiSeq_4R	GGAGTTCAGACGTGTGCTCTTCCGATCTCCTGAGAGAGTTGCAGCAAGC
LacI_MiSeq_5F	CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNGATACGACGATACCGAAGACAG
LacI_MiSeq_5R	GGAGTTCAGACGTGTGCTCTTCCGATCTGAGGCTTTTGACTTTCTGCTAAG
MiSeq_adapter_F_N501	AATGATACGGCGACCACCGAGATCTACACTAGATCGCACACTCTTTCCCTACACGACG CT
MiSeq_adapter_F_N502	AATGATACGGCGACCACCGAGATCTACACCTCTCTATACACTCTTTCCCTACACGACG CT
MiSeq_adapter_F_N503	AATGATACGGCGACCACCGAGATCTACACTATCCTCTACACTCTTTCCCTACACGACG CT
MiSeq_adapter_F_N504	AATGATACGGCGACCACCGAGATCTACACAGAGTAGAACACTCTTTCCCTACACGACG CT
MiSeq_adapter_R_N701	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTGACTGGAGTTCAGACGTGTGCTC
MiSeq_adapter_R_N702	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTGACTGGAGTTCAGACGTGTGCTC
MiSeq_adapter_R_N703	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTGACTGGAGTTCAGACGTGTGCTC
MiSeq_adapter_R_N704	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTGACTGGAGTTCAGACGTGTGCTC
MiSeq_adapter_R_N705	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTGACTGGAGTTCAGACGTGTGCTC
MiSeq_adapter_R_N706	CAAGCAGAAGACGGCATACGAGATCATGCCTAGTGACTGGAGTTCAGACGTGTGCTC
MiSeq_adapter_R_N707	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGTGACTGGAGTTCAGACGTGTGC
MiSeq_adapter_R_N708	CAAGCAGAAGACGGCATACGAGATCCTCTCTGGTGACTGGAGTTCAGACGTGTGC
MiSeq_adapter_R_N709	CAAGCAGAAGACGGCATACGAGATAGCGTAGCGTGACTGGAGTTCAGACGTGTGC
MiSeq_adapter_R_N710	CAAGCAGAAGACGGCATACGAGATCAGCCTCGGTGACTGGAGTTCAGACGTGTGC
MiSeq_adapter_R_N711	CAAGCAGAAGACGGCATACGAGATTGCCTCTTGTGACTGGAGTTCAGACGTGTGC
MiSeq_adapter_R_N712	CAAGCAGAAGACGGCATACGAGATTCCTCTACGTGACTGGAGTTCAGACGTGTGC
Sanger sequencing primers	
Primer name	Sequence
specR_seq_3'	GAAGGATGTCGCTGCCGACT
VR_seq_R	ATTACCGCCTTTGAGTGAGC
pLtetO_seq_3'	TCCCTATCAGTGATAGAGATTGACATCCCT

References

1. Carlson, R. *Biodesic 2011 Bioeconomy Update.* Seattle: Biodesic LLC, 2011. Print. DocID: 20110811_01.

2. Dietrich, J.A., A.E. McKee, and J.D. Keasling. *High-throughput metabolic engineering: advances in small-molecule screening and selection.* Annu. Rev. Biochem. 2010. **79**: p. 563-590.

3. Nakamura, C.E. and G.M. Whited. *Metabolic engineering for the microbial production of 1,3-propanediol.* Curr. Opin. Biotechnol. 2003. **14**(5): p. 454-459.

4. Yim, H., R. Haselbeck, W. Niu, C. Pujol-Baxley, A. Burgard, J. Boldt, J. Khandurina, J.D. Trawick, R.E. Osterhout, R. Stephen, J. Estadilla, S. Teisan, H.B. Schreyer, S. Andrae, T.H. Yang, S.Y. Lee, M.J. Burk, and S. Van Dien. *Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol.* Nat. Chem. Biol. 2011. **7**(7): p. 445-452.

5. Saxena, R.K., P. Anand, S. Saran, and J. Isar. *Microbial production of 1,3-propanediol: Recent developments and emerging opportunities.* Biotechnol. Adv. 2009. **27**(6): p. 895-913.

6. Salis, H.M., E.A. Mirsky, and C.A. Voigt. *Automated design of synthetic ribosome binding sites to control protein expression.* Nat Biotechnol. 2009. **27**(10): p. 946-50.

7. Kosuri, S., D.B. Goodman, G. Cambray, V.K. Mutalik, Y. Gao, A.P. Arkin, D. Endy, and G.M. Church. *Composability of regulatory sequences controlling transcription and translation in Escherichia coli.* Proc Natl Acad Sci U S A. 2013. **110**(34): p. 14024-9.

8. Siegel, J.B., A. Zanghellini, H.M. Lovick, G. Kiss, A.R. Lambert, J.L. St Clair, J.L. Gallaher, D. Hilvert, M.H. Gelb, B.L. Stoddard, K.N. Houk, F.E. Michael, and D. Baker. *Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction.* Science. 2010. **329**(5989): p. 309-313.

9. Gibson, D.G., J.I. Glass, C. Lartigue, V.N. Noskov, R.Y. Chuang, M.A. Algire, G.A. Benders, M.G. Montague, L. Ma, M.M. Moodie, C. Merryman, S. Vashee, R. Krishnakumar, N. Assad-Garcia, C. Andrews-Pfannkoch, E.A. Denisova, L. Young, Z.Q. Qi, T.H. Segall-Shapiro, C.H. Calvey, P.P. Parmar, C.A. Hutchison, 3rd, H.O. Smith, and J.C. Venter. *Creation of a bacterial cell controlled by a chemically synthesized genome.* Science. 2010. **329**(5987): p. 52-6.

10. Caspi, R., T. Altman, R. Billington, K. Dreher, H. Foerster, C.A. Fulcher, T.A. Holland, I.M. Keseler, A. Kothari, A. Kubo, M. Krummenacker, M. Latendresse, L.A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D.S. Weaver, D. Weerasinghe, P. Zhang, and P.D. Karp. *The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases.* Nucleic Acids Res. 2014. **42**: p. D459-71.

11. Lee, J.W., D. Na, J.M. Park, J. Lee, S. Choi, and S.Y. Lee. *Systems metabolic engineering of microorganisms for natural and non-natural chemicals.* Nat Chem Biol. 2012. **8**(6): p. 536-46.

12. Cho, A., H. Yun, J.H. Park, S.Y. Lee, and S. Park. *Prediction of novel synthetic pathways for the production of desired chemicals.* BMC Syst Biol. 2010. **4**: p. 35.

13. Dhamankar, H. and K.L.J. Prather. *Microbial chemical factories: recent advances in pathway engineering for synthesis of value added chemicals.* Curr. Opin. Struct. Biol. 2011. **21**(4): p. 488-494.

14. Blankschien, M.D., J.M. Clomburg, and R. Gonzalez. *Metabolic engineering of Escherichia coli for the production of succinate from glycerol.* Metab Eng. 2010. **12**(5): p. 409-19.

15. Tseng, H.C., C.L. Harwell, C.H. Martin, and K.L. Prather. *Biosynthesis of chiral 3-hydroxyvalerate from single propionate-unrelated carbon sources in metabolically engineered E. coli.* Microb Cell Fact. 2010. **9**: p. 96.

16. Moon, T.S., S.H. Yoon, A.M. Lanza, J.D. Roy-Mayhew, and K.L. Prather. *Production of glucaric acid from a synthetic pathway in recombinant Escherichia coli.* Appl Environ Microbiol. 2009. **75**(3): p. 589-95.

17. Trantas, E., N. Panopoulos, and F. Ververidis. *Metabolic engineering of the complete pathway leading to heterologous biosynthesis of various flavonoids and stilbenoids in Saccharomyces cerevisiae.* Metab Eng. 2009. **11**(6): p. 355-66.

18. Leonard, E., K.-H. Lim, P.-N. Saw, and M.A.G. Koffas. *Engineering central metabolic pathways for high-level flavonoid production in Escherichia coli.* Appl. Environ. Microbiol. 2007. **73**(12): p. 3877-3886.

19. Santos, C.N.S., M. Koffas, and G. Stephanopoulos. *Optimization of a heterologous pathway for the production of flavonoids from glucose.* Metab. Eng. 2011. **13**(4): p. 392-400.

20. Anthony, J.R., L.C. Anthony, F. Nowroozi, G. Kwon, J.D. Newman, and J.D. Keasling. *Optimization of the mevalonate-based isoprenoid biosynthetic pathway in Escherichia coli for production of the anti-malarial drug precursor amorpha-4,11-diene.* Metab Eng. 2009. **11**(1): p. 13-9.

21. Shastri, A. and J. Morgan. *Calculation of theoretical yields in metabolic networks.* Biochem Mol Biol Educ. 2004. **32**(5): p. 314-8.

22. Burgard, A.P., P. Pharkya, and C.D. Maranas. *Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization.* Biotechnol Bioeng. 2003. **84**(6): p. 647-57.

23. Orth, J.D., I. Thiele, and B.Ø. Palsson. *What is flux balance analysis?* Nat. Biotechnol. 2010. **28**(3): p. 245-248.

24. Lewis, N.E., H. Nagarajan, and B.Ø. Palsson. *Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods.* Nat. Rev. Microbiol. 2012. **10**(4): p. 291-305.

25. Bayer, T.S., D.M. Widmaier, K. Temme, E.A. Mirsky, D.V. Santi, and C.A. Voigt. *Synthesis of methyl halides from biomass using engineered microbes.* J Am Chem Soc. 2009. **131**(18): p. 6508-15.

26. Jiang, L., E.A. Althoff, F.R. Clemente, L. Doyle, D. Röthlisberger, A. Zanghellini, J.L. Gallaher, J.L. Betker, F. Tanaka, C.F. Barbas, D. Hilvert, K.N. Houk, B.L. Stoddard, and D. Baker. *De novo computational design of retro-aldol enzymes.* Science. 2008. **319**(5868): p. 1387-1391.

27. Röthlisberger, D., O. Khersonsky, A.M. Wollacott, L. Jiang, J. DeChancie, J. Betker, J.L. Gallaher, E.A. Althoff, A. Zanghellini, O. Dym, S. Albeck, K.N. Houk, D.S. Tawfik, and D. Baker. *Kemp elimination catalysts by computational enzyme design.* Nature. 2008. **453**(7192): p. 190-195.

28. Giger, L., S. Caner, R. Obexer, P. Kast, D. Baker, N. Ban, and D. Hilvert. *Evolution of a designed retro-aldolase leads to complete active site remodeling.* Nat Chem Biol. 2013. **9**(8): p. 494-8.

29. Backus, M.P. and J.F. Stauffer. *The Production and Selection of a Family of Strains in Penicillium chrysogenum.* Mycologia. 1955. **47**(4): p. 429.

30. Wang, H.H., F.J. Isaacs, P.A. Carr, Z.Z. Sun, G. Xu, C.R. Forest, and G.M. Church. *Programming cells by multiplex genome engineering and accelerated evolution.* Nature. 2009. **460**(7257): p. 894-U133.

31. Isaacs, F.J., P.A. Carr, H.H. Wang, M.J. Lajoie, B. Sterling, L. Kraal, A.C. Tolonen, T.A. Gianoulis, D.B. Goodman, N.B. Reppas, C.J. Emig, D. Bang, S.J. Hwang, M.C. Jewett, J.M. Jacobson, and G.M. Church. *Precise manipulation of chromosomes in vivo enables genome-wide codon replacement.* Science. 2011. **333**(6040): p. 348-353.

32. Wang, H.H., H. Kim, L. Cong, J. Jeong, D. Bang, and G.M. Church. *Genome-scale promoter engineering by coselection MAGE.* Nat. Methods. 2012. **9**(6): p. 591-593.

33. Jiang, W., D. Bikard, D. Cox, F. Zhang, and L.A. Marraffini. *RNA-guided editing of bacterial genomes using CRISPR-Cas systems.* Nat. Biotechnol. 2013. **31**(3): p. 233-239.

34. DiCarlo, J.E., J.E. Norville, P. Mali, X. Rios, J. Aach, and G.M. Church. *Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems.* Nucleic Acids Res. 2013. **41**(7): p. 4336-43.

35. Jakociunas, T., I. Bonde, M. Herrgard, S.J. Harrison, M. Kristensen, L.E. Pedersen, M.K. Jensen, and J.D. Keasling. *Multiplex metabolic pathway engineering using CRISPR/Cas9 in Saccharomyces cerevisiae.* Metab Eng. 2015. **28**: p. 213-22.

36. Santos, C.N.S., W. Xiao, and G. Stephanopoulos. *Rational, combinatorial, and genomic approaches for engineering L-tyrosine production in Escherichia coli.* Proc. Natl. Acad. Sci. U.S.A. 2012. **109**(34): p. 13538-13543.

37. Alper, H., J. Moxley, E. Nevoigt, G.R. Fink, and G. Stephanopoulos. *Engineering yeast transcription machinery for improved ethanol tolerance and production.* Science. 2006. **314**(5805): p. 1565-8.

38. Zhao, H., J. Li, B. Han, X. Li, and J. Chen. *Improvement of oxidative stress tolerance in Saccharomyces cerevisiae through global transcription machinery engineering.* J Ind Microbiol Biotechnol. 2014. **41**(5): p. 869-78.

39. Leproust, E.M., B.J. Peck, K. Spirin, H.B. McCuen, B. Moore, E. Namsaraev, and M.H. Caruthers. *Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process.* Nucleic Acids Res. 2010. **38**(8): p. 2522-2540.

40. Kosuri, S., N. Eroshenko, E.M. Leproust, M. Super, J. Way, J.B. Li, and G.M. Church. *Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips.* Nat. Biotechnol. 2010. **28**(12): p. 1295-1299.

41. Carlson, R. *Time for New DNA Synthesis and Sequencing Cost Curves.* Last update 2 Feb 2014. Accessed 3 Mar 2015. Available from: http://www.synthesis.cc/2014/02/time-for-new-cost-curves-2014.html.

42. Zweigenbaum, J., K. Heinig, S. Steinborner, T. Wachs, and J. Henion. *High-Throughput Bioanalytical LC/MS/MS Determination of Benzodiazepines in Human Urine: 1000 Samples per 12 Hours.* Analytical Chemistry. 1999. **71**(13): p. 2294-2300.

43. Raman, S., J.K. Rogers, N.D. Taylor, and G.M. Church. *Evolution-guided optimization of biosynthetic pathways.* Proc Natl Acad Sci U S A. 2014. **111**(50): p. 17803-8.

44. Rogers, J.K., C.D. Guzman, N.D. Taylor, S. Raman, K. Anderson, and G.M. Church. *Synthetic biosensors for precise gene control and real-time monitoring of metabolites.* Nucleic Acids Res. 2015. **43**(15): p. 7648-7660.

45. Lutz, R. and H. Bujard. *Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements.* Nucleic Acids Res. 1997. **25**(6): p. 1203-1210.

46. Swain, P.S., M.B. Elowitz, and E.D. Siggia. *Intrinsic and extrinsic contributions to stochasticity in gene expression.* Proc. Natl. Acad. Sci. U.S.A. 2002. **99**(20): p. 12795-12800.

47. Dietrich, J.A., D.L. Shis, A. Alikhani, and J.D. Keasling. *Transcription factor-based screens and synthetic selections for microbial small-molecule biosynthesis.* ACS Synth Biol. 2013. **2**(1): p. 47-58.

48. Fowler, C.C., E.D. Brown, and Y. Li. *Using a riboswitch sensor to examine coenzyme B(12) metabolism and transport in E. coli.* Chem. Biol. 2010. **17**(7): p. 756-765.

49. van Sint Fiet, S., J.B. van Beilen, and B. Witholt. *Selection of biocatalysts for chemical synthesis.* Proc. Natl. Acad. Sci. U.S.A. 2006. **103**(6): p. 1693-1698.

50. Zheng, J., V. Sagar, A. Smolinsky, C. Bourke, N. LaRonde-LeBlanc, and T.A. Cropp. *Structure and function of the macrolide biosensor protein, MphR(A), with and without erythromycin.* J. Mol. Biol. 2009. **387**(5): p. 1250-1260.

51. Zhang, F.Z., J.M. Carothers, and J.D. Keasling. *Design of a dynamic sensor-regulator system for production of chemicals and fuels derived from fatty acids.* Nature Biotechnology. 2012. **30**(4): p. 354-U166.

52. Weickert, M.J. and S. Adhya. *Isorepressor of the gal regulon in Escherichia coli.* J. Mol. Biol. 1992. **226**(1): p. 69-83.

53. Monterrubio, R., L. Baldoma, N. Obradors, J. Aguilar, and J. Badia. *A common regulator for the operons encoding the enzymes involved in D-galactarate, D-glucarate, and D-glycerate utilization in Escherichia coli.* J Bacteriol. 2000. **182**(9): p. 2672-4.

54. Lewis, M., G. Chang, N.C. Horton, M.A. Kercher, H.C. Pace, M.A. Schumacher, R.G. Brennan, and P. Lu. *Crystal structure of the lactose operon repressor and its complexes with DNA and inducer.* Science. 1996. **271**(5253): p. 1247-1254.

55. Schell, M.A. *Molecular biology of the LysR family of transcriptional regulators.* Annu. Rev. Microbiol. 1993. **47**: p. 597-626.

56. Craven, S.H., O.C. Ezezika, S. Haddad, R.A. Hall, C. Momany, and E.L. Neidle. *Inducer responses of BenM, a LysR-type transcriptional regulator from Acinetobacter baylyi ADP1.* Mol. Microbiol. 2009. **72**(4): p. 881-894.

57. Siedler, S., G. Schendzielorz, S. Binder, L. Eggeling, S. Bringer, and M. Bott. *SoxR as a single-cell biosensor for NADPH-consuming enzymes in Escherichia coli.* ACS Synth Biol. 2014. **3**(1): p. 41-7.

58. Terán, W., A. Felipe, A. Segura, A. Rojas, J.-L. Ramos, and M.T. Gallegos. *Antibiotic-dependent induction of Pseudomonas putida DOT-T1E TtgABC efflux pump is mediated by the drug binding repressor TtgR.* Antimicrob. Agents Chemother. 2003. **47**(10): p. 3067-3072.

59. Canosa, I., L. Yuste, and F. Rojo. *Role of the alternative sigma factor sigmaS in expression of the AlkS regulator of the Pseudomonas oleovorans alkane degradation pathway.* J. Bacteriol. 1999. **181**(6): p. 1748-1754.

60. Mauzy, C.A. and M.A. Hermodson. *Structural and functional analyses of the repressor, RbsR, of the ribose operon of Escherichia coli.* Protein Sci. 1992. **1**(7): p. 831-42.

61. Garmendia, J., D. Devos, A. Valencia, and V. de Lorenzo. *A la carte transcriptional regulators: unlocking responses of the prokaryotic enhancer-binding protein XylR to non-natural effectors.* Mol. Microbiol. 2001. **42**(1): p. 47-59.

62. Teo, W.S. and M.W. Chang. *Bacterial XylRs and synthetic promoters function as genetically encoded xylose biosensors in Saccharomyces cerevisiae.* Biotechnol J. 2015. **10**(2): p. 315-22.

63. Collins, C.H., F.H. Arnold, and J.R. Leadbetter. *Directed evolution of Vibrio fischeri LuxR for increased sensitivity to a broad spectrum of acyl-homoserine lactones.* Mol. Microbiol. 2005. **55**(3): p. 712-723.

64. Collins, C.H., J.R. Leadbetter, and F.H. Arnold. *Dual selection enhances the signaling specificity of a variant of the quorum-sensing transcriptional activator LuxR.* Nat. Biotechnol. 2006. **24**(6): p. 708-712.

65. Looger, L.L., M.A. Dwyer, J.J. Smith, and H.W. Hellinga. *Computational design of receptor and sensor proteins with novel functions.* Nature. 2003. **423**(6936): p. 185-90.

66. Tang, S.Y. and P.C. Cirino. *Design and application of a mevalonate-responsive regulatory protein.* Angew Chem Int Ed Engl. 2011. **50**(5): p. 1084-6.

67. Feng, J., B.W. Jester, C.E. Tinberg, D.J. Mandell, M.S. Antunes, R. Chari, K.J. Morey, X. Rios, J.I. Medford, G.M. Church, S. Fields, and D. Baker. *A general strategy to construct small molecule biosensors in eukaryotes.* Elife. 2015. **4**.

68. San Martin, A., S. Ceballo, F. Baeza-Lehnert, R. Lerchundi, R. Valdebenito, Y. Contreras-Baeza, K. Alegria, and L.F. Barros. *Imaging mitochondrial flux in single cells with a FRET sensor for pyruvate.* PLoS One. 2014. **9**(1): p. e85780.

69. Maynard-Smith, L.A., L.-C. Chen, L.A. Banaszynski, A.G.L. Ooi, and T.J. Wandless. *A directed approach for engineering conditional protein stability using biologically silent small molecules.* J. Biol. Chem. 2007. **282**(34): p. 24866-24872.

70. Lynch, S.A. and J.P. Gallivan. *A flow cytometry-based screen for synthetic riboswitches.* Nucleic Acids Res. 2009. **37**(1): p. 184-192.

71. Muranaka, N., K. Abe, and Y. Yokobayashi. *Mechanism-guided library design and dual genetic selection of synthetic OFF riboswitches.* Chembiochem. 2009. **10**(14): p. 2375-81.

72. Muranaka, N., V. Sharma, Y. Nomura, and Y. Yokobayashi. *An efficient platform for genetic selection and screening of gene switches in Escherichia coli.* Nucleic Acids Res. 2009. **37**(5): p. e39.

73. Peroza, E.A., J.C. Ewald, G. Parakkal, J.M. Skotheim, and N. Zamboni. *A genetically encoded Forster resonance energy transfer sensor for monitoring in vivo trehalose-6-phosphate dynamics.* Anal Biochem. 2015. **474**: p. 1-7.

74. Tang, S.Y., S. Qian, O. Akinterinwa, C.S. Frei, J.A. Gredell, and P.C. Cirino. *Screening for enhanced triacetic acid lactone production by recombinant Escherichia coli expressing a designed triacetic acid lactone reporter.* J Am Chem Soc. 2013. **135**(27): p. 10099-103.

75. de Los Santos, E.L., J.T. Meyerowitz, S.L. Mayo, and R.M. Murray. *Engineering Transcriptional Regulator Effector Specificity Using Computational Design and In Vitro Rapid Prototyping: Developing a Vanillin Sensor.* ACS Synth Biol. 2015. [Epub ahead of print]

76. Vinkenborg, J.L., T.J. Nicolson, E.A. Bellomo, M.S. Koay, G.A. Rutter, and M. Merkx. *Genetically encoded FRET sensors to monitor intracellular Zn2+ homeostasis.* Nat Methods. 2009. **6**(10): p. 737-40.

77. Jha, R.K., S. Chakraborti, T.L. Kern, D.T. Fox, and C.E. Strauss. *Rosetta comparative modeling for library design: Engineering alternative inducer specificity in a transcription factor.* Proteins. 2015. [Epub ahead of print]

78. Swint-Kruse, L. and K.S. Matthews. *Allostery in the LacI/GalR family: variations on a theme.* Current Opinion in Microbiology. 2009. **12**(2): p. 129-137.

79. Gossen, M., S. Freundlieb, G. Bender, G. Müller, W. Hillen, and H. Bujard. *Transcriptional activation by tetracyclines in mammalian cells.* Science. 1995. **268**(5218): p. 1766-1769.

80. Licitra, E.J. and J.O. Liu. *A three-hybrid system for detecting small ligand-protein receptor interactions.* Proc. Natl. Acad. Sci. U.S.A. 1996. **93**(23): p. 12817-12821.

81. Chien, C.T., P.L. Bartel, R. Sternglanz, and S. Fields. *The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest.* Proceedings of the National Academy of Sciences. 1991. **88**(21): p. 9578-9582.

82. Werstuck, G. and M.R. Green. *Controlling gene expression in living cells through small molecule-RNA interactions.* Science. 1998. **282**(5387): p. 296-298.

83. Wittmann, A. and B. Suess. *Engineered riboswitches: Expanding researchers' toolbox with synthetic RNA regulators.* FEBS Lett. 2012. **586**(15): p. 2076-83.

84. Serganov, A. and E. Nudler. *A decade of riboswitches.* Cell. 2013. **152**(1-2): p. 17-24.

85. Breaker, R.R. *Prospects for riboswitch discovery and analysis.* Mol. Cell. 2011. **43**(6): p. 867-879.

86. Berens, C. and B. Suess. *Riboswitch engineering - making the all-important second and third steps.* Curr Opin Biotechnol. 2015. **31**: p. 10-5.

87. Morris, M.C. *Fluorescent biosensors - probing protein kinase function in cancer and drug discovery.* Biochim. Biophys. Acta. 2013. **1834**(7): p. 1387-1395.

88. Capra, E.J. and M.T. Laub. *Evolution of two-component signal transduction systems.* Annu. Rev. Microbiol. 2012. **66**: p. 325-347.

89. Wang, B., M. Barahona, and M. Buck. *A modular cell-based biosensor using engineered genetic logic circuits to detect and integrate multiple environmental signals.* Biosens Bioelectron. 2013. **40**(1): p. 368-76.

90. Schloss, P.D. and J. Handelsman. *Biotechnological prospects from metagenomics.* Curr. Opin. Biotechnol. 2003. **14**(3): p. 303-310.

91. Wilson, M.C. and J. Piel. *Metagenomic approaches for exploiting uncultivated bacteria as a resource for novel biosynthetic enzymology.* Chem. Biol. 2013. **20**(5): p. 636-647.

92. Tinberg, C.E., S.D. Khare, J. Dou, L. Doyle, J.W. Nelson, A. Schena, W. Jankowski, C.G. Kalodimos, K. Johnsson, B.L. Stoddard, and D. Baker. *Computational design of ligand-binding proteins with high affinity and selectivity.* Nature. 2013. **501**(7466): p. 212-6.

93. Taylor, N.D., A.S. Garruss, R. Moretti, S. Chan, M.A. Arbing, D. Cascio, J.K. Rogers, F.J. Isaacs, S. Kosuri, D. Baker, S. Fields, G.M. Church, and S. Raman. *Engineering an allosteric transcription factor to respond to new ligands.* Nat Methods. 2015. [Epub ahead of print]

94. Farmer, W.R. and J.C. Liao. *Improving lycopene production in Escherichia coli by engineering metabolic control.* Nat Biotechnol. 2000. **18**(5): p. 533-7.

95. Binder, S., G. Schendzielorz, N. Stabler, K. Krumbach, K. Hoffmann, M. Bott, and L. Eggeling. *A high-throughput approach to identify genomic variants of bacterial metabolite producers at the single-cell level.* Genome Biol. 2012. **13**(5): p. 40.

96. Mohrle, V., M. Stadler, and G. Eberz. *Biosensor-guided screening for macrolides.* Anal Bioanal Chem. 2007. **388**(5-6): p. 1117-25.

97. Behzadian, F., H. Barjeste, S. Hosseinkhani, and A.R. Zarei. *Construction and characterization of Escherichia coli whole-cell biosensors for toluene and related compounds.* Curr Microbiol. 2011. **62**(2): p. 690-6.

98. Holtz, W.J. and J.D. Keasling. *Engineering static and dynamic control of synthetic pathways.* Cell. 2010. **140**(1): p. 19-23.

99. Xu, P., L. Li, F. Zhang, G. Stephanopoulos, and M. Koffas. *Improving fatty acids production by engineering dynamic pathway regulation and metabolic control.* Proc Natl Acad Sci U S A. 2014. **111**(31): p. 11299-304.

100. Liu, D., Y. Xiao, B.S. Evans, and F. Zhang. *Negative feedback regulation of fatty acid production based on a malonyl-CoA sensor-actuator.* ACS Synth Biol. 2015. **4**(2): p. 132-40.

101. Scalcinati, G., C. Knuf, S. Partow, Y. Chen, J. Maury, M. Schalk, L. Daviet, J. Nielsen, and V. Siewers. *Dynamic control of gene expression in Saccharomyces cerevisiae engineered for the production of plant sesquitepene alpha-santalene in a fed-batch mode.* Metab Eng. 2012. **14**(2): p. 91-103.

102. Dahl, R.H., F. Zhang, J. Alonso-Gutierrez, E. Baidoo, T.S. Batth, A.M. Redding-Johanson, C.J. Petzold, A. Mukhopadhyay, T.S. Lee, P.D. Adams, and J.D. Keasling. *Engineering dynamic pathway regulation using stress-response promoters.* Nat Biotechnol. 2013. **31**(11): p. 1039-46.

103. Hong, S.H., M. Hegde, J. Kim, X. Wang, A. Jayaraman, and T.K. Wood. *Synthetic quorum-sensing circuit to control consortial biofilm formation and dispersal in a microfluidic device.* Nat Commun. 2012. **3**: p. 613.

104. Lütke-Eversloh, T. and G. Stephanopoulos. *Feedback inhibition of chorismate mutase/prephenate dehydrogenase (TyrA) of Escherichia coli: generation and characterization of tyrosine-insensitive mutants.* Appl. Environ. Microbiol. 2005. **71**(11): p. 7224-7228.

105. Yang, J., S.W. Seo, S. Jang, S.I. Shin, C.H. Lim, T.Y. Roh, and G.Y. Jung. *Synthetic RNA devices to expedite the evolution of metabolite-producing microbes.* Nat Commun. 2013. **4**: p. 1413.

106. Gallegos, M.T., R. Schleif, A. Bairoch, K. Hofmann, and J.L. Ramos. *Arac/XylS family of transcriptional regulators.* Microbiol. Mol. Biol. Rev. 1997. **61**(4): p. 393-410.

107. Ramos, J.L., M. Martínez-Bueno, A.J. Molina-Henares, W. Terán, K. Watanabe, X. Zhang, M.T. Gallegos, R. Brennan, and R. Tobes. *The TetR family of transcriptional repressors.* Microbiol. Mol. Biol. Rev. 2005. **69**(2): p. 326-356.

108. Tropel, D. and J.R. van der Meer. *Bacterial transcriptional regulators for degradation pathways of aromatic compounds.* Microbiol Mol Biol Rev. 2004. **68**(3): p. 474-500.

109. Wang, K.H., E.S.C. Oakes, R.T. Sauer, and T.A. Baker. *Tuning the strength of a bacterial N-end rule degradation signal.* J. Biol. Chem. 2008. **283**(36): p. 24600-24607.

110. Chen, H., M. Bjerknes, R. Kumar, and E. Jay. *Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of Escherichia coli mRNAs.* Nucleic Acids Res. 1994. **22**(23): p. 4953-4957.

111. DeVito, J.A. *Recombineering with tolC as a Selectable/Counter-selectable Marker: remodeling the rRNA Operons of Escherichia coli.* Nucleic Acids Research. 2008. **36**(1): p. e4.

112. Xu, P., S. Ranganathan, Z.L. Fowler, C.D. Maranas, and M.A.G. Koffas. *Genome-scale metabolic network modeling results in minimal interventions that cooperatively force carbon flux towards malonyl-CoA.* Metab. Eng. 2011. **13**(5): p. 578-587.

113. Takai, A., R. Nishi, Y. Joe, and H. Ito, *L-tyrosine-producing bacterium and a method for producing L-tyrosine*. 2005. Ajinomoto Co., Inc. European Patent EP1616940 B1.

114. Burlingame, R. and P.J. Chapman. *Catabolism of phenylpropionic acid and its 3-hydroxy derivative by Escherichia coli.* J. Bacteriol. 1983. **155**(1): p. 113-121.

115. Díaz, E., A. Ferrández, and J.L. García. *Characterization of the hca cluster encoding the dioxygenolytic pathway for initial catabolism of 3-phenylpropionic acid in Escherichia coli K-12.* J. Bacteriol. 1998. **180**(11): p. 2915-2923.

116. Gehring, A.M., K.A. Bradley, and C.T. Walsh. *Enterobactin biosynthesis in Escherichia coli: isochorismate lyase (EntB) is a bifunctional enzyme that is phosphopantetheinylated by EntD and then acylated by EntE using ATP and 2,3-dihydroxybenzoate.* Biochemistry. 1997. **36**(28): p. 8495-8503.

117. Yoon, S.-H., T.S. Moon, P. Iranpour, A.M. Lanza, and K.J. Prather. *Cloning and characterization of uronate dehydrogenases from two pseudomonads and Agrobacterium tumefaciens strain C58.* J. Bacteriol. 2009. **191**(5): p. 1565-1573.

118. Shiue, E. and K.L.J. Prather. *Improving d-glucaric acid production from myo-inositol in E. coli by increasing MIOX stability and myo-inositol transport.* Metab. Eng. 2013. **22C**: p. 22-31.

119. Moon, T.S., J.E. Dueber, E. Shiue, and K.L.J. Prather. *Use of modular, synthetic scaffolds for improved production of glucaric acid in engineered E. coli.* Metab. Eng. 2010. **12**(3): p. 298-305.

120. Canonaco, F., T.A. Hess, S. Heri, T. Wang, T. Szyperski, and U. Sauer. *Metabolic flux response to phosphoglucose isomerase knock-out in Escherichia coli and impact of overexpression of the soluble transhydrogenase UdhA.* FEMS Microbiol. Lett. 2001. **204**(2): p. 247-252.

121. Fraenkel, D.G. *The accumulation of glucose 6-phosphate from glucose and its effect in an Escherichia coli mutant lacking phosphoglucose isomerase and glucose 6-phosphate dehydrogenase.* J. Biol. Chem. 1968. **243**(24): p. 6451-6457.

122. Grant, A.W., G. Steel, H. Waugh, and E.M. Ellis. *A novel aldo-keto reductase from Escherichia coli can increase resistance to methylglyoxal toxicity.* FEMS Microbiol. Lett. 2003. **218**(1): p. 93-99.

123. Carr, P.A. and G.M. Church. *Genome engineering.* Nat. Biotechnol. 2009. **27**(12): p. 1151-1162.

124. Nyerges, Á., B. Csorgő, I. Nagy, D. Latinovics, B. Szamecz, G. Pósfai, and C. Pál. *Conditional DNA repair mutants enable highly precise genome engineering.* Nucleic Acids Res. 2014. **42**(8): p. e62.

125. Andersen, J.B., C. Sternberg, L.K. Poulsen, S.P. Bjorn, M. Givskov, and S. Molin. *New unstable variants of green fluorescent protein for studies of transient gene expression in bacteria.* Appl. Environ. Microbiol. 1998. **64**(6): p. 2240-2246.

126. Datsenko, K.A. and B.L. Wanner. *One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products.* Proc. Natl. Acad. Sci. U.S.A. 2000. **97**(12): p. 6640-6645.

127. Lu, T.K., A.S. Khalil, and J.J. Collins. *Next-generation synthetic gene networks.* Nat. Biotechnol. 2009. **27**(12): p. 1139-1150.

128. Milo-Landesman, D., M. Surana, I. Berkovich, A. Compagni, G. Christofori, N. Fleischer, and S. Efrat. *Correction of hyperglycemia in diabetic mice transplanted with reversibly immortalized pancreatic beta cells controlled by the tet-on regulatory system.* Cell Transplant. 2001. **10**(7): p. 645-650.

129. Zhang, J., M.K. Jensen, and J.D. Keasling. *Development of biosensors and their application in metabolic engineering.* Curr Opin Chem Biol. 2015. **28**: p. 1-8.

130. Süel, G.M., S.W. Lockless, M.A. Wall, and R. Ranganathan. *Evolutionarily conserved networks of residues mediate allosteric communication in proteins.* Nat Struct Biol. 2002. **10**(1): p. 59-69.

131. Markiewicz, P., L.G. Kleina, C. Cruz, S. Ehret, and J.H. Miller. *Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence.* J. Mol. Biol. 1994. **240**(5): p. 421-433.

132. Suckow, J., P. Markiewicz, L.G. Kleina, J. Miller, B. Kisters-Woike, and B. Müller-Hill. *Genetic studies of the Lac repressor. XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure.* J. Mol. Biol. 1996. **261**(4): p. 509-523.

133. Raman, S., N. Taylor, N. Genuth, S. Fields, and G.M. Church. *Engineering allostery.* Trends Genet. 2014. **30**(12): p. 521-528.

134. Fazelinia, H., P.C. Cirino, and C.D. Maranas. *Extending Iterative Protein Redesign and Optimization (IPRO) in protein library design for ligand specificity.* Biophys. J. 2007. **92**(6): p. 2120-2130.

135. Tang, S.-Y., H. Fazelinia, and P.C. Cirino. *AraC regulatory protein mutants with altered effector specificity.* J. Am. Chem. Soc. 2008. **130**(15): p. 5267-5271.

136. Galvão, T.C., M. Mencía, and V. de Lorenzo. *Emergence of novel functions in transcriptional regulators by regression to stem protein types.* Mol. Microbiol. 2007. **65**(4): p. 907-919.

137. AbuOun, M., P.F. Suthers, G.I. Jones, B.R. Carter, M.P. Saunders, C.D. Maranas, M.J. Woodward, and M.F. Anjum. *Genome scale reconstruction of a Salmonella metabolic model: comparison of similarity and differences with a commensal Escherichia coli strain.* J. Biol. Chem. 2009. **284**(43): p. 29480-29488.

138. Swint-Kruse, L., C.R. Elam, J.W. Lin, D.R. Wycuff, and K. Shive Matthews. *Plasticity of quaternary structure: twenty-two ways to form a LacI dimer.* Protein Sci. 2001. **10**(2): p. 262-276.

139. Swint-Kruse, L., H. Zhan, B.M. Fairbanks, A. Maheshwari, and K.S. Matthews. *Perturbation from a Distance: Mutations that Alter LacI Function through Long-Range Effects †.* Biochemistry. 2003. **42**(47): p. 14004-14016.

140. Xu, J. and K.S. Matthews. *Flexibility in the Inducer Binding Region Is Crucial for Allostery in the Escherichia coliLactose Repressor.* Biochemistry. 2009. **48**(22): p. 4988-4998.

141. Mirny, L.A. and M.S. Gelfand. *Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors*. J. Mol. Biol. 2002. **321**(1): p. 7-20.

142. Kalinina, O.V., A.A. Mironov, M.S. Gelfand, and A.B. Rakhmaninova. *Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families.* Protein Sci. 2004. **13**(2): p. 443-456.

143. Pei, J., W. Cai, L.N. Kinch, and N.V. Grishin. *Prediction of functional specificity determinants from protein sequences using log-likelihood ratios.* Bioinformatics. 2006. **22**(2): p. 164-171.

144. Bell, C.E. and M. Lewis. *A closer view of the conformation of the Lac repressor bound to operator.* Nat Struct Biol. 2000. **7**(3): p. 209-214.

145. Stemmer, W.P. *DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution.* Proc. Natl. Acad. Sci. U.S.A. 1994. **91**(22): p. 10747-10751.

146. Guntas, G., T.J. Mansell, J.R. Kim, and M. Ostermeier. *Directed evolution of protein switches and their application to the creation of ligand-binding proteins.* Proc. Natl. Acad. Sci. U.S.A. 2005. **102**(32): p. 11224-11229.

147. Qin, Y., Y. Zhang, H. He, J. Zhu, G. Chen, W. Li, and Z. Liang. *Screening and identification of a fungal β-glucosidase and the enzymatic synthesis of gentiooligosaccharide.* Appl. Biochem. Biotechnol. 2011. **163**(8): p. 1012-1019.

148. Gossen, M. and H. Bujard. *Tight control of gene expression in mammalian cells by tetracycline-responsive promoters.* Proc. Natl. Acad. Sci. U.S.A. 1992. **89**(12): p. 5547-5551.

149. Pédelacq, J.-D., S. Cabantous, T. Tran, T.C. Terwilliger, and G.S. Waldo. *Engineering and characterization of a superfolder green fluorescent protein.* Nat. Biotechnol. 2006. **24**(1): p. 79-88.

150. McCoy, A.J., R.W. Grosse-Kunstleve, P.D. Adams, M.D. Winn, L.C. Storoni, and R.J. Read. *Phaser crystallographic software.* J Appl Crystallogr. 2007. **40**(Pt 4): p. 658-674.

151. Hawkins, P.C.D. and A. Nicholls. *Conformer generation with OMEGA: learning from the data set and the analysis of failures.* J Chem Inf Model. 2012. **52**(11): p. 2919-2936.

152. Kabsch, W. *XDS.* Acta Crystallogr. D Biol. Crystallogr. 2010. **66**(Pt 2): p. 125-132.

153. Strong, M., M.R. Sawaya, S. Wang, M. Phillips, D. Cascio, and D. Eisenberg. *Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from Mycobacterium tuberculosis.* Proc. Natl. Acad. Sci. U.S.A. 2006. **103**(21): p. 8060-8065.

154. Emsley, P., B. Lohkamp, W.G. Scott, and K. Cowtan. *Features and development of Coot.* Acta Crystallogr. D Biol. Crystallogr. 2010. **66**(Pt 4): p. 486-501.

155. Murshudov, G.N., A.A. Vagin, and E.J. Dodson. *Refinement of macromolecular structures by the maximum-likelihood method.* Acta Crystallogr. D Biol. Crystallogr. 1997. **53**(Pt 3): p. 240-255.

156. Winn, M.D., M.N. Isupov, and G.N. Murshudov. *Use of TLS parameters to model anisotropic displacements in macromolecular refinement.* Acta Crystallogr. D Biol. Crystallogr. 2001. **57**(Pt 1): p. 122-133.

157. Altschul, S.F., T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res. 1997. **25**(17): p. 3389-3402.

158. Larkin, M.A., G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, and D.G. Higgins. *Clustal W and Clustal X version 2.0.* Bioinformatics. 2007. **23**(21): p. 2947-2948.

159. Majumdar, A., S. Rudikoff, and S. Adhya. *Purification and properties of Gal repressor:pL-galR fusion in pKC31 plasmid vector.* J. Biol. Chem. 1987. **262**(5): p. 2326-2331.

160. Meinhardt, S., M.W. Manley, N.A. Becker, J.A. Hessman, L.J. Maher, and L. Swint-Kruse. *Novel insights from hybrid LacI/GalR proteins: family-wide functional attributes and biologically significant variation in transcription repression.* Nucleic Acids Res. 2012. **40**(21): p. 11139-11154.

161. Magoč, T. and S.L. Salzberg. *FLASH: fast length adjustment of short reads to improve genome assemblies.* Bioinformatics. 2011. **27**(21): p. 2957-2963.

162. Kent, W.J. *BLAT--the BLAST-like alignment tool.* Genome Res. 2002. **12**(4): p. 656-664.

163. Bolstad, B.M., R.A. Irizarry, M. Astrand, and T.P. Speed. *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics. 2003. **19**(2): p. 185-193.

164. Wickham, H., *ggplot2: elegant graphics for data analysis*. 2009, New York: Springer.

165. Rodriguez, G.J., R. Yao, O. Lichtarge, and T.G. Wensel. *Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors.* Proc. Natl. Acad. Sci. U.S.A. 2010. **107**(17): p. 7787-7792.

166. Bai, F., R.W. Branch, D.V. Nicolau, T. Pilizota, B.C. Steel, P.K. Maini, and R.M. Berry. *Conformational spread as a mechanism for cooperativity in the bacterial flagellar switch.* Science. 2010. **327**(5966): p. 685-689.

167. Hilser, V.J. *An Ensemble View of Allostery.* Science. 2010. **327**(5966): p. 653-654.

168. Goodey, N.M. and S.J. Benkovic. *Allosteric regulation and catalysis emerge via a common route.* Nat. Chem. Biol. 2008. **4**(8): p. 474-482.

169. Amaro, R.E., A. Sethi, R.S. Myers, V.J. Davisson, and Z.A. Luthey-Schulten. *A network of conserved interactions regulates the allosteric signal in a glutamine amidotransferase.* Biochemistry. 2007. **46**(8): p. 2156-2173.

170. Gandhi, P.S., Z. Chen, F.S. Mathews, and E. Di Cera. *Structural identification of the pathway of long-range communication in an allosteric enzyme.* Proc. Natl. Acad. Sci. U.S.A. 2008. **105**(6): p. 1832-1837.

171. Ricketson, D., U. Hostick, L. Fang, K.R. Yamamoto, and B.D. Darimont. *A conformational switch in the ligand-binding domain regulates the dependence of the glucocorticoid receptor on Hsp90.* J. Mol. Biol. 2007. **368**(3): p. 729-741.

172. Masterson, L.R., A. Mascioni, N.J. Traaseth, S.S. Taylor, and G. Veglia. *Allosteric cooperativity in protein kinase A.* Proc. Natl. Acad. Sci. U.S.A. 2008. **105**(2): p. 506-511.

173. Lewis, M. *The lac repressor.* Comptes Rendus Biologies. 2005. **328**(6): p. 521-548.

174. Kleina, L.G. and J.H. Miller. *Genetic studies of the lac repressor. XIII. Extensive amino acid replacements generated by the use of natural and synthetic nonsense suppressors.* J. Mol. Biol. 1990. **212**(2): p. 295-318.

175. Müller-Hartmann, H. and B. Müller-Hill. *The side-chain of the amino acid residue in position 110 of the Lac repressor influences its allosteric equilibrium.* J. Mol. Biol. 1996. **257**(3): p. 473-478.

176. de Juan, D., F. Pazos, and A. Valencia. *Emerging methods in protein co-evolution.* Nature Publishing Group. 2013. **14**(4): p. 249-261.

177. Parente, D.J. and L. Swint-Kruse. *Multiple Co-Evolutionary Networks Are Supported by the Common Tertiary Scaffold of the LacI/GalR Proteins.* PLoS ONE. 2013. **8**(12): p. e84398.

178. Shulman, A.I., C. Larson, D.J. Mangelsdorf, and R. Ranganathan. *Structural determinants of allosteric ligand activation in RXR heterodimers.* Cell. 2004. **116**(3): p. 417-429.

179. Fowler, D.M., C.L. Araya, S.J. Fleishman, E.H. Kellogg, J.J. Stephany, D. Baker, and S. Fields. *High-resolution mapping of protein sequence-function relationships.* Nat. Methods. 2010. **7**(9): p. 741-746.

180. Fowler, D.M. and S. Fields. *Deep mutational scanning: a new style of protein science.* Nat. Methods. 2014. **11**(8): p. 801-807.

181. Fujino, Y., R. Fujita, K. Wada, K. Fujishige, T. Kanamori, L. Hunt, Y. Shimizu, and T. Ueda. *Robust in vitro affinity maturation strategy based on interface-focused high-throughput mutational scanning.* Biochem. Biophys. Res. Commun. 2012. **428**(3): p. 395-400.

182. Whitehead, T.A., A. Chevalier, Y. Song, C. Dreyfus, S.J. Fleishman, C. De Mattos, C.A. Myers, H. Kamisetty, P. Blair, I.A. Wilson, and D. Baker. *Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing.* Nat. Biotechnol. 2012. **30**(6): p. 543-548.

183. Ernst, A., D. Gfeller, Z. Kan, S. Seshagiri, P.M. Kim, G.D. Bader, and S.S. Sidhu. *Coevolution of PDZ domain–ligand interactions analyzed by high-throughput phage display and deep sequencing.* Mol. BioSyst. 2010. **6**(10): p. 1782.

184. Araya, C.L., D.M. Fowler, W. Chen, I. Muniez, J.W. Kelly, and S. Fields. *A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function.* Proc. Natl. Acad. Sci. U.S.A. 2012. **109**(42): p. 16858-16863.

185. Starita, L.M., J.N. Pruneda, R.S. Lo, D.M. Fowler, H.J. Kim, J.B. Hiatt, J. Shendure, P.S. Brzovic, S. Fields, and R.E. Klevit. *Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis.* Proc. Natl. Acad. Sci. U.S.A. 2013. **110**(14): p. E1263-72.

186. Findlay, G.M., E.A. Boyle, R.J. Hause, J.C. Klein, and J. Shendure. *Saturation editing of genomic regions by multiplex homology-directed repair.* Nature. 2014.

187. Cheng, A.A., H. Ding, and T.K. Lu. *Enhanced killing of antibiotic-resistant bacteria enabled by massively parallel combinatorial genetics.* Proc. Natl. Acad. Sci. U.S.A. 2014. **111**(34): p. 12462-12467.

188. McLaughlin, J., Richard N, F.J. Poelwijk, A. Raman, W.S. Gosal, and R. Ranganathan. *The spatial architecture of protein function and adaptation.* Nature. 2013. **490**(7422): p. 138-142.

189. Daily, M.D., T.J. Upadhyaya, and J.J. Gray. *Contact rearrangements form coupled networks from local motions in allosteric proteins.* Proteins. 2008. **71**(1): p. 455-466.

190. Fukami-Kobayashi, K., Y. Tateno, and K. Nishikawa. *Parallel evolution of ligand specificity between LacI/GalR family repressors and periplasmic sugar-binding proteins.* Mol. Biol. Evol. 2003. **20**(2): p. 267-277.

191. Maeder, M.L., S. Thibodeau-Beganny, A. Osiak, D.A. Wright, R.M. Anthony, M. Eichtinger, T. Jiang, J.E. Foley, R.J. Winfrey, J.A. Townsend, E. Unger-Wallace, J.D. Sander, F. Muller-Lerch, F. Fu, J. Pearlberg, C. Gobel, J.P. Dassie, S.M. Pruett-Miller, M.H. Porteus, D.C. Sgroi, A.J. Iafrate, D. Dobbs, P.B. McCray, Jr., T. Cathomen, D.F. Voytas, and J.K. Joung. *Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification.* Mol Cell. 2008. **31**(2): p. 294-301.

192. Bourguet, W., P. Germain, and H. Gronemeyer. *Nuclear receptor ligand-binding domains: three-dimensional structures, molecular interactions and pharmacological implications.* Trends Pharmacol. Sci. 2000. **21**(10): p. 381-388.

193. Gronemeyer, H., J.-A. Gustafsson, and V. Laudet. *Principles for modulation of the nuclear receptor superfamily.* Nat Rev Drug Discov. 2004. **3**(11): p. 950-964.

194. Wurtz, J.M., W. Bourguet, J.P. Renaud, V. Vivat, P. Chambon, D. Moras, and H. Gronemeyer. *A canonical structure for the ligand-binding domain of nuclear receptors.* Nat Struct Biol. 1996. **3**(1): p. 87-94.

195. Galperin, M.Y. *Structural classification of bacterial response regulators: diversity of output domains and domain combinations.* J. Bacteriol. 2006. **188**(12): p. 4169-4182.

196. Capra, E.J., B.S. Perchuk, E.A. Lubin, O. Ashenberg, J.M. Skerker, and M.T. Laub. *Systematic dissection and trajectory-scanning mutagenesis of the molecular interface that ensures specificity of two-component signaling pathways.* PLoS Genet. 2010. **6**(11): p. e1001220.

197. Ferguson, S.S., W.E. Downey, A.M. Colapietro, L.S. Barak, L. Ménard, and M.G. Caron. *Role of beta-arrestin in mediating agonist-promoted G protein-coupled receptor internalization.* Science. 1996. **271**(5247): p. 363-366.

198. Schlessinger, J. *Cell signaling by receptor tyrosine kinases.* Cell. 2000. **103**(2): p. 211-225.

199. Stynen, B., H. Tournu, J. Tavernier, and P. Van Dijck. *Diversity in genetic in vivo methods for protein-protein interaction studies: from the yeast two-hybrid system to the mammalian split-luciferase system.* Microbiol. Mol. Biol. Rev. 2012. **76**(2): p. 331-382.

200. Petschnigg, J., B. Groisman, M. Kotlyar, M. Taipale, Y. Zheng, C.F. Kurat, A. Sayad, J.R. Sierra, M. Mattiazzi Usaj, J. Snider, A. Nachman, I. Krykbaeva, M.S. Tsao, J. Moffat, T. Pawson, S. Lindquist, I. Jurisica, and I. Stagljar. *The mammalian-membrane two-hybrid assay (MaMTH) for probing membrane-protein interactions in human cells.* Nat Methods. 2014. **11**(5): p. 585-92.

201. Kittanakom, S., M. Barrios-Rodiles, J. Petschnigg, A. Arnoldo, V. Wong, M. Kotlyar, L.E. Heisler, I. Jurisica, J.L. Wrana, C. Nislow, and I. Stagljar. *CHIP-MYTH: A novel interactive proteomics method for the assessment of agonist-dependent interactions of the human β2-adrenergic receptor.* Biochem. Biophys. Res. Commun. 2014. **445**(4): p. 746-756.

202. Yan, Y.-X., D.M. Boldt-Houle, B.P. Tillotson, M.A. Gee, B.J. D'Eon, X.-J. Chang, C.E.M. Olesen, and M.A.J. Palmer. *Cell-based high-throughput screening assay system for monitoring G protein-coupled receptor activation using beta-galactosidase enzyme complementation technology.* J Biomol Screen. 2002. **7**(5): p. 451-459.

203. Escobedo, J.A., S. Navankasattusas, W.M. Kavanaugh, D. Milfay, V.A. Fried, and L.T. Williams. *cDNA cloning of a Novel 85 kd protein that has SH2 domains and regulates binding of PI3-kinase to the PDGF β-receptor.* Cell. 1991. **65**(1): p. 75-82.

204. González-Vera, J.A. *Probing the kinome in real time with fluorescent peptides.* Chem Soc Rev. 2012. **41**(5): p. 1652-1664.

205. Chen, C.-A., R.-H. Yeh, X. Yan, and D.S. Lawrence. *Biosensors of protein kinase action: from in vitro assays to living cells.* Biochim. Biophys. Acta. 2004. **1697**(1-2): p. 39-51.

206. Baird, G.S., D.A. Zacharias, and R.Y. Tsien. *Circular permutation and receptor insertion within green fluorescent proteins.* Proc. Natl. Acad. Sci. U.S.A. 1999. **96**(20): p. 11241-11246.

207. Goodman, D.B., G.M. Church, and S. Kosuri. *Causes and effects of N-terminal codon bias in bacterial genes.* Science. 2013. **342**(6157): p. 475-9.

208. Das, R. and D. Baker. *Macromolecular modeling with rosetta.* Annu. Rev. Biochem. 2008. **77**: p. 363-382.

209. Bilitchenko, L., A. Liu, S. Cheung, E. Weeding, B. Xia, M. Leguia, J.C. Anderson, and D. Densmore. *Eugene--a domain specific language for specifying and constraining synthetic biological parts, devices, and systems.* PLoS One. 2011. **6**(4): p. e18882.

210. Daber, R., S. Stayrook, A. Rosenberg, and M. Lewis. *Structural analysis of lac repressor bound to allosteric effectors.* J. Mol. Biol. 2007. **370**(4): p. 609-619.

211. Nivón, L.G., R. Moretti, and D. Baker. *A Pareto-optimal refinement method for protein design scaffolds.* PLoS ONE. 2013. **8**(4): p. e59004.

212. O'Boyle, N.M., M. Banck, C.A. James, C. Morley, T. Vandermeersch, and G.R. Hutchison. *Open Babel: An open chemical toolbox.* J Cheminform. 2011. **3**: p. 33.

213. Fleishman, S.J., A. Leaver-Fay, J.E. Corn, E.-M. Strauch, S.D. Khare, N. Koga, J. Ashworth, P. Murphy, F. Richter, G. Lemmon, J. Meiler, and D. Baker. *RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite.* PLoS ONE. 2011. **6**(6): p. e20161.

214. Lazaridis, T. and M. Karplus. *Effective energy function for proteins in solution.* Proteins. 1999. **35**(2): p. 133-152.

215. Thyme, S.B., D. Baker, and P. Bradley. *Improved modeling of side-chain--base interactions and plasticity in protein--DNA interface design.* J. Mol. Biol. 2012. **419**(3-4): p. 255-274.